ED 230 559	TM 820 491
AUTHOR	Kingston, Neal M.; Dorans, Neil J.
TITLE	The Feasibility of Using Item Response Theory as a Psychometric Model for the GRE Aptitude Test.
INSTITUTION	Educational Testing Service, Princeton, N.J.; Graduate Record Examinations Board, Princeton, N.J.
REPORT NO	ETS-RR-82-12; GREB-79-12P
PUB DATE	Apr 82
NOTE	168p.; Some tables may be marginally legible due to small print.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	Aptitude Tests; *Graduate Study; Higher Education; *Latent Trait Theory; *Mathematical Models; Psychometrics; Standardized Tests; *Statistical Analysis; *Testing Programs; *Test Items
IDENTIFIERS	*Graduate Record Examinations; Robustness; Three Parameter Model

ABSTRACT

The feasibility of using item response theory (IRT) as a psychometric model for the Graduate Record Examination (GRE) Aptitude Test was addressed by assessing the reasonableness of the assumptions of item response theory for GRE item types and examinee populations. Items from four forms and four administrations of the GRE Aptitude Test were calibrated using the three-parameter logistic item response model. Three equating methods were compared in this research: equipercentile equating, linear equating, and item response theory true score equating. Various data collection designs (for both IRT and non-IRT methods) and several item parameter linking procedures (for the IRT equatings) were employed. The IRT methods produced quantitative scaled score means and standard deviations that were higher and lower, respectively, than those produced by the linear and equipercentile methods. The most notable finding in the analytical equatings was the sensitivity of the precalibration design (used only for the IRT equating method) to practice effects on analytical items, particularly for the analysis of explanations item type. Since the precalibration design is the data collection method most appealing (for administrative reasons) for equating the GRE Aptitude Test in a test disclosure environment, this sensitivity might present a problem for any equating method. (PN)

THE FEASIBILITY OF USING ITEM RESPONSE THEORY AS A PSYCHOMETRIC MODEL FOR THE GRE APTITUDE TEST

Neal M. Kingston

and

Neil J. Dorans

GRE Board Professional Report GREB No. 79-12P

ETS Research Report 82-12

April 1982

This report presents the findings of a research project funded by and carried out under the auspices of the Graduate Record Examinations Board.

GRE BOARD RESEARCH REPORTS
FOR GENERAL AUDIENCE

Altman, R. A. and Wallmark, M. M. A Summary of Data from the Graduate Programs and Admissions Manual. GREB No. 74-1R, January 1975.

Baird, L. L. An Inventory of Documented Accomplishments. GREB No. 77-3R, June 1979.

Baird, L. L. Cooperative Student Survey (The Graduates [$2.50 each], and Careers and Curricula). GREB No. 70-4R, March 1973.

Baird, L. L. The Relationship Between Ratings of Graduate Departments and Faculty Publication Rates. GREB No. 77-2aR, November 1980.

Baird, L. L. and Knapp, J. E. The Inventory of Documented Accomplishments for Graduate Admissions: Results of a Field Trial Study of Its Reliability, Short-Term Correlares, and Evaluation. GREB No. 78-3R, August 1981.

Burns, R. L. Graduate Admissions and Fellowship Selection Policies and Procedures (Part I and II). GREB No. 69-5R, July 1970.

Centra, J. A. How Universities Evaluate Faculty Performance: A Survey of Department Heads. GREB No. 75-5bR, July 1977. ($1.50 each)

Centra, J. A. Women, Men and the Doctorate. GREB No. 71-10R, September 1974. ($3.50 each)

Clark, M. J. The Assessment of Quality in Ph.D. Programs: A Preliminary Report on Judgments by Graduate Deans. GREB No. 72-7aR, October 1974.

Clark, M. J. Program Review Practices of University Departments. GREB No. 75-5aR, July 1977. ($1.00 each)

Devore, R. and McPeek, M. A Study of the Content of Three GRE Advanced Tests. GREB No. 78-4R, March 1982.

Donlon, T. F. Annotated Bibliography of Test Speededness. GREB No. 76-9R, June 1979.

Flaugher, R. L. The New Definitions of Test Fairness In Selection: Developments and Implications. GREB No. 72-4R, May 1974.

Fortna, R. O. Annotated Bibliography of the Graduate Record Examinations. July 1979.

Frederiksen, N. and Ward, W. C. Measures for the Study of Creativity in Scientific Problem-Solving. May 1978.

Hartnett, R. T. Sex Differences in the Environments of Graduate Students and Faculty. GREB No. 77-2bR, March 1981.

Hartnett, R. T. The Information Needs of Prospective Graduate Students. GREB No. 77-8R, October 1979.

Hartnett, R. T. and Willingham, W. W. The Criterion Problem: What Measure of Success in Graduate Education? GREB No. 77-4R, March 1979.

Knapp, J. and Hamilton, I. B. The Effect of Nonstandard Undergraduate Assessment and Reporting Practices on the Graduate School Admissions Process. GREB No. 76-14R, July 1978.

Lannholm, G. V. and Parry, M. E. Programs for Disadvantaged Students in Graduate Schools. GREB No. 69-1R, January 1970.

Miller, R. and Wild, C. L. Restructuring the Graduate Record Examinations Aptitude Test. GRE Board Technical Report, June 1979.

Reilly, R. R. Critical Incidents of Graduate Student Performance. GREB No. 70-5R, June 1974.

Rock, D., Werts, C. An Analysis of Time Related Score Increments and/or Decrements for GRE Repeaters across Ability and Sex Groups. GREB No. 77-9R, April 1979.

Rock, D. A. The Prediction of Doctorate Attainment in Psychology, Mathematics and Chemistry. GREB No. 69-6aR, June 1974.

Schrader, W. B. GRE Scores as Predictors of Career Achievement in History. GREB No. 76-1bR, November 1980.

Schrader, W. B. Admissions Test Scores as Predictors of Career Achievement in Psychology. GREB No. 76-1aR, September 1978.

Swinton, S. S. and Powers, D. E. A Study of the Effects of Special Preparation on GRE Analytical Scores and Item Types. GREB No. 78-2R, January 1982.

Wild, C. L. Summary of Research on Restructuring the Graduate Record Examinations Aptitude Test. February 1979.

Wild, C. L. and Durso, R. Effect of Increased Test-Taking Time on Test Scores by Ethnic Group, Age, and Sex. GREB No. 76-6R, June 1979.

Wilson, K. M. The GRE Cooperative Validity Studies Project. GREB No. 75-8R, June 1979.

Wiltsey, R. G. Doctoral Use of Foreign Languages: A Survey. GREB No. 70-14R, 1972. (Highlights $1.00, Part I $2.00, Part II $1.50).

Witkin, H. A.; Moore, C. A.; Oltman, P. K., Goodenough, D. R.; Friedman, F.; and Owen, D. R. A Longitudinal Study of the Role of Cognitive Styles in Academic Evolution During the College Years. GREB No. 76-10R, February 1977 ($5.00 each).

THE FEASIBILITY OF USING ITEM RESPONSE THEORY AS A PSYCHOMETRIC MODEL FOR THE GRE APTITUDE TEST

Neal M. Kingston

and

Neil J. Dorans

GRE Board Professional Report GREB No. 79-12P

April 1982

Abstract

The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test was addressed by assessing the reasonableness of the assumptions of item response theory for GRE item types and examinee populations. Items from four forms and four administrations of the GRE Aptitude Test were calibrated using the three-parameter logistic item response model (one form was given at two administrations and one administration used two forms; the exact relationships between forms and administrations are given in Test Forms and Populations section of this report).

The unidimensionality assumption of item response theory was addressed in a variety of ways. Previous factor analytic research on the GRE Aptitude Test was reviewed to assess the dimensionality of the test and to extract information pertinent to the construction of sets of homogeneous items. On the basis of this review, separate calibrations of discrete verbal items and reading comprehension items were run, in addition to calibrations on all verbal items, because two strong dimensions on the verbal scale were identified in the factor analytic research.

Local independence of item responses is a consequence of the unidimensionality assumption. To test the weak form of the local independence condition, partial correlations, both with and without a correction for guessing, among items with ability partialled out were computed and factor analyzed. Violations of local independence were observed in both verbal item types and quantitative item types. These violations were basically consistent with expectations based on the factor analytic review.

Fit of the three-parameter logistic model to GRE Aptitude Test data was assessed by comparing estimated item-ability regressions, i.e., item response functions, with empirical item-ability regressions. The three-parameter model fit all verbal item types reasonably well. The fit to data interpretation items, regular math items, analytical reasoning items, and logical diagrams items also seemed acceptable. The model fit quantitative comparison items least well. The analysis of explanations item type was also not fit well by the three-parameter logistic model.

The stability of item parameter estimates for different samples was assessed. Item difficulty estimates exhibited a large degree of stability, followed by item discrimination parameter estimates. The hard-to-estimate lower asymptote or pseudoguessing parameter exhibited the least temporal stability.

The sensitivity of item parameter estimates to the lack of unidimensionality that produced the local independence violations was examined. The discrete verbal and all verbal calibrations of discrete verbal items produced more similiar estimates of item discrimination than the reading comprehension and all verbal calibrations of reading comprehension items, reflecting the larger correlations that overall verbal ability estimates had with discrete verbal ability estimates. As compared to item discrimination estimates, item difficulty estimates exhibited much less sensitivity to homogeneity of item sets. The estimates of the lower asymptote were, for the most part, fairly robust to homogeneity of item calibration set.

The comparability of ability estimates based on homogeneous item sets (reading comprehension items or discrete verbal items) with estimates based on all verbal items was examined. Correlations among overall verbal ability estimates, discrete verbal ability estimates, and reading comprehension ability estimates provided evidence for the existence of two distinct, highly correlated verbal abilities that can be combined to produce a composite ability that resembles the overall verbal ability defined by the calibration of all verbal items together.

Three equating methods were compared in this research: equipercentile equating, linear equating, and item response theory true score equating. Various data collection designs (for both IRT and non-IRT methods) and several item parameter linking procedures (for the IRT equatings) were employed. The equipercentile and linear equatings of the verbal scales were more similar to each other than they were to the IRT equatings. The degree of similarity among the scaled score distributions produced by the various equating methods, data collection designs, and linking procedures was greater for the verbal equatings than for either the quantitative or analytical equatings. In almost every comparison, the IRT methods produced quantitative scaled score means and standard deviations that were higher and lower, respectively, than those produced by the linear and equipercentile methods. The most notable finding in the analytical equatings was the sensitivity of the precalibration design (in this study, used only for the IRT equating method) to practice effects on analytical items, particularly for the analysis of explanations item type. Since the precalibration design is the data collection method most appealing (for administrative reasons) for equating the GRE Aptitude Test in a test disclosure environment, this sensitivity might present a problem for any equating method.

In sum, the item response theory model and IRT true score equating, using the precalibration data collection design, appear most applicable to the verbal section, less applicable to the quantitative section because of possible dimensionality problems with data interpretation items and instances of nonmontonicity for the quantitative comparison items, and least applicable to the analytical section because of severe practice effects associated with the analysis of explanations item type. Expected revisions of the analytical section, particularly the removal of the troublesome analysis of explanations item type, should enhance the fit and applicability of the three-parameter model to the analytical section. Planned revisions of the verbal section should not substantially affect the satisfactory fit of the model to verbal item types. The heterogeneous quantitative section might present problems for item response theory. It must be remembered, however, that these same (and other) factors that affect IRT based equatings may also affect other equating methods.

TABLE OF CONTENTS

INTRODUCTION

The use of item response theory as a psychometric model for the GRE Aptitude Test can provide a powerful set of statistical tools for analysis of items and tests, maintenance of score scales via equating, and development of better and more efficient test forms (Cowell, 1979; Hambleton and Cook, 1977; Hambleton, 1980; Lord, 1977, 1980a; Marco, 1977; and Warm, 1978). Determination of the applicability of IRT methods to the GRE Aptitude Test requires an assessment of the psychometric feasibility of using IRT as a mathematical model for item responses on the GRE Aptitude Test. Psychometric feasibility can be addressed by examining the reasonableness and importance of the underlying assumptions of IRT for GRE populations and item types. The present research addresses the reasonableness of these assumptions and the robustness of IRT methods to violations of these assumptions.

Assumptions of Item Response Theory

Item response theory provides a mathematical expression for the probability of success on an item as a function of a single characteristic of the individual answering the item, his or her ability, and multiple characteristics of the item. This mathematical expression is called an item response function. Both on psychometric grounds and for reasons of tractability, a reasonable mathematical form for the item response function of a multiple choice item is the three-parameter logistic model,

$$(1)\quad P_g(\theta) = c_g + \frac{1 - c_g}{1 + e^{-1.7\, a_g\, (\theta - b_g)}} ,$$

where

$P_g(\theta)$ is the probability that an examinee with ability θ answers item g correctly,

e is the base of the system of natural logarithms approximately equal to 2.7183,

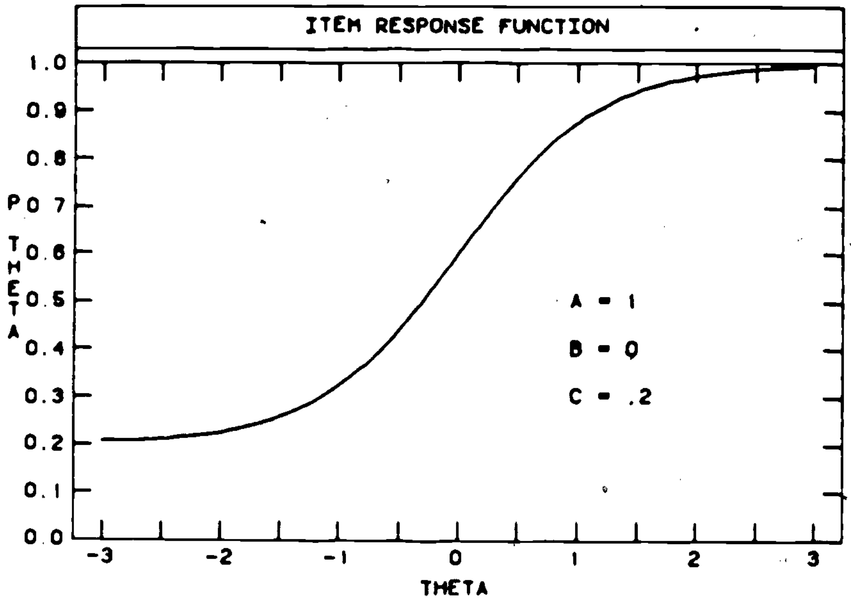
a_g is a measure of item discrimination for item g,

b_g is a measure of item difficulty for item g, and

c_g is the lower asymptote of the item response curve, the probability of very low ability examinees answering item g correctly.

In equation (1), θ is the ability parameter, a characteristic of the examinee, and a_g, b_g and c_g are item parameters that determine the shape of the item response function (see Figure 1).

Figure 1



One of the major assumptions of IRT embodied in equation (1) is that the set of items under study is unidimensional, i.e., the probability of successful response by examinees to a set of items can be modelled with only one ability parameter, θ. The second major assumption embodied in equation (1) is that the probability of successful performance on an item can be adequately described by the three-parameter logistic model.

One consequence of the unidimensionality assumption is the mathematical concept of local independence. There are two forms of local independence, weak and strong. The strong form can be stated as:

$$(2) \qquad \text{Prob } (\underline{V} = \underline{v} \P \theta) = \prod_{g=1}^{n} P_g(\theta)^{u_g} Q_g(\theta)^{1-u_g} \text{ , where}$$

$\underline{V}$ is a vector random variable of binary responses (right or wrong) for the n items,

$\underline{v}$ is a particular vector response pattern,

θ is the ability level,

u_g is an examinee's binary response to item g, either 1 or 0,

$P_g(\theta)$ is the probability of a correct response for an examinee of ability θ,

$Q_g(\theta)$ is $1 - P_g(\theta)$, the probability of an incorrect response for an examinee of ability θ, and

n is the number of items on the test.

This form is equivalent to saying that, at each ability level, item responses are statistically independent. The weak form of local independence states that at each θ, item responses are uncorrelated.

Assessing the Reasonableness of the Assumptions

A major purpose of the present research is to assess the reasonableness of the assumptions of IRT for GRE item types and populations. There is wide agreement (Bejar, 1980; Hambleton, Swaminathan, Cook, Eignor, & Gifford, 1978; Lord, 1980a) that no single method exists for conclusively determining whether a set of responses to a set of items is unidimensional. Consequently, a variety of approaches were employed to assess the dimensionality assumption.

Review of pertinent factor analytic research. Four factor analytic research studies conducted on the GRE Aptitude Test were reviewed in order to assess the dimensionality of the test and to extract information pertinent to the construction of sets of homogeneous items. These studies were also examined to extract hypotheses about the GRE Aptitude Test that could be tested at later stages of the research.

Weak form of local independence. As stated earlier, local independence among items is a mathematical consequence of the unidimensionality assumption. If responses to a set of items are unidimensional, these responses are statistically independent at a given level of ability. The local independence condition was tested by computing $r_{gh\cdot\theta}$, the tetrachoric correlation between items g and h with estimated θ partialled out (Warm, 1978, p. 101), for every pair of items in sets of apparently homogeneous items. These correlations were computed both with and without a correction for guessing (Carroll, 1945). The partial correlations were examined to identify items with large positive correlations. The matrices of the partial correlations were then factor analyzed. Results of this semi-nonlinear factor analysis were compared with previous linear factor analytic results. Hypotheses were generated to explain these results.

Item-ability regressions. The item response function obtained from the estimated item parameters can be viewed as an estimation of the theoretical form for the regression of item score (1 = a correct response, 0 = an incorrect response) onto underlying ability. In other words, the item response function describes expected item performance as a function of ability. Actual item performance for a given estimated ability level was obtained from the data and plotted for various levels of ability to approximate an empirical item-ability regression (Hambleton, 1980; Stocking, 1980). Visual inspection of how closely the estimated item-ability regression captured the empirical item-ability regression provided information about how well the three-parameter logistic model fit the data. Comparison of item-ability regressions for items calibrated in both homogeneous sets (e.g., all reading comprehension items) and heterogeneous sets (e.g., all verbal items) was of particular interest.

Comparisons based on homogeneous and heterogeneous subsets of items. In addition to visual inspection of the estimated and empirical item-ability regressions, examination of the comparability of item parameter estimates was used to assess the effects of heterogeneity on the fit of the logistic model. Correlations between item parameter estimates for the same items calibrated in a homogeneous set and in a heterogeneous set were computed to index the degree of similiarity between the item-ability regressions. Mean differences between item parameter estimates also provided information about the relative fit of the logistic model for sets of homogeneous and heterogeneous items.

Position or practice effect. The unidimensionality assumption implies that the only systematic influences on item performance are the individual's ability and characteristics of the item. Given knowledge of an individual's ability, knowledge about that individual's performance on one item does not add any information for forecasting that individual's performance on another item. In other words, since ability and item characteristics are the only systematic influences on item performance, knowledge of that individual's performance on other items is superfluous. One practical consequence of the unidimensionality assumption is that item position should have no effect on item performance because, if item position affected item performance, then something other than ability would be having a systematic effect on item performance. In short, if there is a position effect or practice effect on item performance, the unidimensionality assumption is violated. In the present research, the same items appeared in two different locations on two forms of the GRE Aptitude test, enabling us to ascertain whether a position effect existed.

Practice effect, though a problem stemming from data collection design, can have a major impact on the equating of test forms. Practice effect can occur when items appear in the second section of the same item type. Also, a general effect, perhaps induced by fatigue, might occur on any items appearing late in a test. Any such systematic bias might not appear when the item was later used in another position in an operational section of the test, which would contribute to an incorrect equating. This problem will exist (though not necessarily to the same extent) with any equating method that makes use of data collected in one portion of the test to equate scores based on a different portion of the test.

This report examines the impact of practice effect on IRT true score equating. Practice effects are analyzed in greater detail in another research report (Kingston & Dorans, 1982).

Robustness of IRT Equating to Violations of Assumptions

Few mathematical models ever fit the data completely. The three-parameter logistic model will not completely explain expected item performance on the GRE Aptitude Test any more than "...'a heavy point swinging without friction on a weightless string' (which) never existed in the real world, but at a certain stage of the process of knowledge ... is a very useful model of a pendulum" (Rasch, 1960). The various methods of assessing the fit of the model described in the section on reasonableness of IRT assumptions provided us with knowledge about the degree to which the model fits the data. This knowledge is synthesized with the results of the equatings in the last section of this report.

REVIEW OF FACTOR ANALYTIC RESEARCH

Four factor analytic research studies conducted on the GRE Aptitude Test were reviewed in order to assess the dimensionality of the test and to extract information pertinent to the construction of sets of homogeneous items. The four studies are:

I. Powers, D. E., Swinton, S. S., & Carlson, A. B. A factor analytic study of the GRE Aptitude Test, GRE Board Professional Report GREB No 75-11P, September 1977.

II. Powers, D. E., Swinton, S. S., Thayer, D., & Yates, A., A factor analytic investigation of seven experimental analytical item types, GRE Board Professional Report GREB No 77-1P, June 1978.

III. Swinton, S. S., & Powers, D. E. A factor analytic study of the restructured GRE Aptitude Test, GRE Board Professional Report GREB No 77-6P, February 1980.

IV. Rock, D. A., Werts, C., & Grandy, J. Construct validity of the GRE across populations---an empirical confirmatory study. Draft Report, 1980.

The first three studies involved factor analyses conducted at the item level on interitem tetrachoric correlations; the third study also involved a factor analysis at the level of item parcels, i.e., items grouped together on the basis of item difficulty and nominal item type, e.g., analogies. Joreskog's (1978) confirmatory factor analysis model was used in the fourth study where the factoring was conducted on correlations among nominal item type parcels.

Study I

The stated purposes of the Powers, Swinton, and Carlson (1977) study were to determine the factor structure of the preanalytical GRE Aptitude Test and to determine the structure of several experimental tests by relating each of these tests to the structure of the operational GRE Aptitude Test. At that time, the operational GRE Aptitude Test was given in three separately timed sections:

I.	Discrete verbal (25 minutes)	(55 items)
	- analogies	(18 items)
	- antonyms or opposites	(20 items)
	- sentence completions	(17 items)
II.	Reading comprehension (50 minutes)	(40 items)
III.	Quantitative (75 minutes)	(55 items)
	- regular math	(40 items)
	- data interpretation	(15 items)

The experimental tests were composed of either reading comprehension items, regular math items, data interpretation items, or quantitative comparison items, which at that time was an experimental item type.

Powers et al. (1977) identified three global factors, one associated with each section of the test: general quantitative ability, general verbal ability or reading comprehension, and vocabulary or discrete verbal ability. In addition, they identified smaller factors including a data interpretation factor, speed factors, and a technical reading comprehension factor.

They used Dwyer (1937) extension analyses to extend factors from the space of the operational GRE Aptitude Test into the space of the experimental items and then examined residuals. They found that the quantitative comparison items were better explained by the general quantitative ability factor than were the data interpretation items already in the quantitative section. In addition, they found that the experimental scientific or technical reading comprehension items were not well explained by the two global verbal ability factors of reading comprehension and vocabulary.

Study II

The stated purposes of the Powers, Swinton, Thayer, and Yates (1978) study were to assess, from a factor analytic point of view, the relationships between two preanalytical versions of the GRE Aptitude Test and seven experimental abstract reasoning or analytical item types and to replicate the factor structure uncovered by Powers et al. (1977).

They identified three global factors on the operational GRE Aptitude Test: general quantitative ability, reading comprehension or connected discourse, and vocabulary or discrete verbal ability. In addition they noted some smaller factors including a data interpretation factor, speed factors on the verbal sections, and a specific content reading comprehension factor. The results of Dwyer extension analyses of these operational factors into the space of each type of analytical item revealed that the logical diagrams and analytical reasoning items tended to load more on the quantitative factors than did the analysis of explanations items, which appeared to be the most complex of these three types of analytical items.

Study III

Since the GRE IRT feasibility research was conducted on the current restructured version of the GRE Aptitude Test, the recently completed Swinton and Powers (1980) factor analysis of the restructured GRE Aptitude Test is the most pertinent of the four factor analytic studies. Forms ZGR1 and ZGR2, the first forms containing analtyical items on an operational basis, were studied by Swinton and Powers to provide a factor analytic description of the new restructured test and to compare this structure to the factor structure of the former test. There are four separately timed operational sections of the restructured GRE Aptitude Test:

I.	Verbal ability (50 minutes)	(80 items)
	- discrete verbal	(55 items)
	- reading comprehension	(25 items)
II.	Quantitative ability (50 minutes)	(55 items)
	- quantitative comparison	(30 items)
	- data interpretation & regular math	(25 items)
III.	Analytical ability (25 minutes)	(40 items)
	- analysis of explanations	(40 items)
IV.	Analytical ability (25 minutes)	(30 items)
	- logical diagrams	
	- analytical reasoning	

Both item level analyses and analyses based on item parcels were performed. First, Swinton and Powers (1980) factored analytical items alone and identified, after a varimax rotation (Kaiser, 1958), six factors: one logical diagrams factor, three analysis of explanations factors, a speed factor, and an analytical reasoning factor. Visual inspection of a plot of eigenvalues from analytical item tetrachoric correlation matrices with communality estimates in the diagonal reveals that Swinton and Powers may have overfactored. On the basis of these plots, it appears that one, maybe two, factors would have been sufficient for the purpose of describing the major dimensions of the analytical section.

Next, Swinton and Powers factored the reduced tetrachoric correlation matrix for all items together and identified four major factors: reading comprehension or general verbal ability, vocabulary or discrete verbal ability, difficult quantitative and easy quantitative. In addition, they identified four smaller factors: a data interpretation factor, a technical reading comprehension factor, and two factors dealing with analytical items. Again, from visual inspection of the eigenvalue plots, it would appear that only four factors are needed to represent the important dimensions of the test.

On the basis of these item level analyses, item parcels were constructed using nominal item type, item difficulty, and in some cases, e.g., the analysis of explanations items, item response key as facets. For example, the 20 antonym items were clustered into five unique parcels composed of four items each and these five item parcels differed in difficulty. A total of 53 item parcels were constructed. The purpose of constructing item parcels is to avoid some of the problems associated with the factoring of binary data, such as the appearance of item difficulty factors and the instability of tetrachorics. Constructing parcels that differed in mean difficulty, however, may have defeated one purpose of constructing the parcels.

In their factor analysis of the 53 item parcels, Swinton and Powers' varimax factors were called verbal reasoning, quantitative, and vocabulary, while the remaining three varimax factors were called technical reading comprehension, data interpretation, and analytical. The six oblimin (Jennrich & Sampson, 1966) factors were called easy items, quantitative, vocabulary, technical reading comprehension, data interpretation, and analytical. Easy items is obviously a difficulty factor. Finally, the geoplane (Yates, 1974) solution produced a reading comprehension and sentence completion factor, a general quantitative factor, a vocabulary factor, an analytical factor, a data interpretation and technical reading comprehension factor, and an easy quantitative factor.

Study IV

To assess the construct validity of the restructured GRE Aptitude Test, Rock, Werts, and Grandy (1980) employed Joreskog's (1978) confirmatory factor analysis model to evaluate various psychometric models for the GRE Aptitude Test by testing progressively more restrictive hypotheses about the relationships between observed scores and underlying true scores or factor scores. Their analysis was performed at the level of nominal item type; 20 scores were produced, odd-even half scores for each of the 10 nominal items types. Since nominal item type score was the level of analysis, their report does not have direct implications for the evaluation of the dimensionality of items. The report, however, is indirectly relevant.

In particular, examination of the 20-by-20 correlation matrix for these 20 odd-even item type scores is informative. The discrete verbal or vocabulary scores all correlate highly. The two reading comprehension, the two quantitative comparisons, and the six analytical all correlate highly. The two data interpretation scores tend to have the lowest correlations with all other scores.

Synthesis

The difficulty factor problem. Before synthesizing these four studies and discussing their implications for GRE IRT equating research, a brief discussion of the perils of using factor analytic techniques with binary data is appropriate.

The common factor model (Thurstone, 1947) is frequently employed to assess the dimensionality of a test or set of tests. It is a model that postulates a linear relationship between observed attributes, such as those measured by tests, and underlying basic attributes or factors.

The appearance of "difficulty factors" complicates the application of factor analytic techniques to binary data such as multiple-choice items. The difficulty factor problem has long been recognized in the psychometric literature. McDonald (1967) presents a brief review of the difficulty factor literature, mentioning work by Guilford (1941), Ferguson

(1941), Wherry and Gaylord (1944), Carroll (1945), Gourlay (1951), and Gibson (1959, 1960) among others. Guilford obtained a factor that was related to item difficulty in his analyses of the Seashore Test of Pitch Discriminations. Ferguson demonstrated that a matrix of phi coefficients for homogeneous items, i.e., items measuring the same ability, would have a rank greater than one if items differed widely in difficulty. Wherry and Gaylord concluded that the appearance of Ferguson's difficulty factor was due to use of the wrong correlation coefficient and recommended use of the tetrachoric correlation for factoring binary data. Both Carroll and Gourlay indicated conditions under which tetrachorics might yield a difficulty factor. Carroll demonstrated that, under guessing conditions, the obtained correlation of tests or items decreases as the tests or items become less similar in difficulty and that the obtained correlation between pairs of items decreases as their average difficulty becomes greater. Gibson claimed that difficulty factors can be considered caused by the nonlinear regression of tests on factors. The point of this brief review is to demonstrate that difficulty factors are a problem to contend with when interpreting the results of factor analytic studies.

Difficulty factors appeared in the Swinton and Powers (1980) factor analysis of the restructured GRE Aptitude Test. The varimax rotation of the unrotated factor matrix, obtained from factoring the reduced tetrachoric correlation matrix among all items, produced a difficult quantitative factor and an easy quantitative factor. On the basis of these results, the authors used item difficulty as a facet in the construction of item parcels. As a consequence, difficulty factors appeared in both the oblimin and geoplane solutions. The appearance of these difficulty factors complicates the interpretation of the results. For the purpose of constructing sets of homogeneous items for the present research, it seemed reasonable to ignore these difficulty factors since the three-parameter logistic IRT model allows for differential difficulty among items.

Implications for GRE IRT feasibility research. Despite the interpretative complications induced by the appearance of difficulty factors, the Swinton and Powers study of the restructured GRE Aptitude Test had definite implications for the construction of sets of homogeneous items for the GRE IRT equating research. Along with the other three factor analytic studies, this study provided strong evidence for the existence of three large global factors: general quantitative ability, reading comprehension or general verbal reasoning, and vocabulary. An obvious implication of this finding is that separation of reading comprehension items from other verbal items would produce two sets of items that are more homogeneous than the original set of all verbal items.

Swinton and Powers provided evidence for the multidimensionality of the analytical scale. They retained six factors for orthogonal rotation. While perhaps six factors are necessary to explain the bulk of the score variance, it is likely that only one or two of these factors represent major psychological dimensions. The factor analysis of the item parcels supports this parsimonious position because it produced a single analytical factor despite the fact that item response choice was one of the facets used in the construction of item parcels. If there are two analytical factors, one is probably a quantitative factor and the other is probably a verbal factor. Examination of the rotated factor patterns revealed that the analytical items loaded highly on both the quantitative and verbal factors.

The identification of a single small analytical factor in the factor analysis of item parcels suggested that separation of analytical item types into more homogeneous sets is unnecessary. On the other hand, the fact that the analytical items loaded highly on the quantitative and verbal factors, particularly reading comprehension, suggested that these items are complex. Unfortunately, the fact that the items load on both quantitative and verbal factors would have made it difficult to construct sets of more homogeneous items. In light of this difficulty and the fact that the composition of the analytical section was under revision, a decision was made to focus on the quantitative and verbal sections and to ignore the analytical section for the most part.

The factor analysis review suggests the existence of a small data interpretation factor and a small technical reading comprehension factor, as well as speed factors, particularly in the verbal section. For the sake of homogeneity, separating data interpretation items from other quantitative items might have been a wise course of action. The same argument could be made for the technical reading comprehension items.

It was decided, however, not to construct separate technical reading comprehension and data interpretation scales as there would not have been enough items in the anchor tests to permit stable linkings of ability scales through item difficulty parameters. For example, it would have been necessary to use an anchor test containing 10 items to link the data interpretation scale for form ZGR1 to the data interpretation scale for form 3CGR1. Outliers could have a large impact on the equation that links these two scales. If the guidelines for score equating pertain to linking of scale through IRT item difficulty parameters, the anchor test should contain a minimum of 20 items.

Another reason for not constructing separate technical reading comprehension and data interpretation scales was the existance of a certain skepticism concerning the importance of these factors. Since both these factors are small, one or both might be tiny minor factors (in the Tucker, Koopman, and Linn (1969) sense) that have been elevated to the level of common factors by overfactoring. In the Tucker, Koopman, and Linn model, a distinction is made between two systematic sources of covariation among observed scores: major factors and minor factors. Major factors are the common factors of the common factor model, systematic sources of covariation among observed scores that are viewed as important psychological dimensions. In contrast, minor factors are systematic sources of covariation among observed scores that also exist in the data but are not a part of the common factor model. These minor factors, which influence performance, are not viewed as important dimensions but rather as nuisance components that negatively affect the fit of the factor model to the data. In an effort to describe all systematic covariation among item scores, Swinton and Powers (1980) may have extracted both major and minor factors.

Technical reading comprehension is possibly a form specific minor factor, dependent upon the unusualness of the particular vocabulary employed in the technical reading passages. On the other hand, data interpretation could well be a unique form of quantitative ability. Both factors may raise interesting questions for future restructuring of the GRE. For the purpose of the present research, however, the small numbers of both data interpretation and technical reading comprehension items precluded construction of separate scales for equating.

The existence of these small minor factors must be kept in mind when comparing the results of IRT equating with conventional linear or equipercentile equating. When confronted with two-dimensional data in which one dimension dominates the other, LOGIST is "drawn toward" the larger dimension as it progresses through its iterative parameter estimation process (Reckase, 1979). Hence, the existence of a small data interpretation factor on the quantitative scale could introduce a discrepancy between IRT equating and conventional linear or equipercentile equating of the quantitative scale because of a differential effect of the data interpretation factor on these two equatings. The data interpretation factor will influence the direction of the quantitative true score dimension and the extent of this influence will depend upon the size of this factor. While LOGIST may ignore this factor and iterate toward a general quantitative dimension, conventional equatings will use the intact true score dimension that is partially influenced by this minor factor. Hence, on a priori grounds we expected a discrepancy between conventional and IRT equatings due to this differential effect of the data interpretation factor. Inspection of the fit of the IRT model to the data interpretation items was expected to provide evidence pertaining to the reasonableness of this hypothesis.

The preceding discussion about the potential differential effect of the data interpretation factor on conventional and IRT equatings has implications for the potential effect of the small technical reading comprehension factor on the comparison of IRT and conventional equatings. This small factor could also induce a discrepancy between the conventional and IRT equatings. Here too, inspection of the fit of the IRT model to the technical reading comprehension items was expected to shed light on the reasonableness of this differential impact hypothesis.

The speed component of the verbal section is a nuisance factor that might complicate comparisons of the results of the IRT equating with the conventional linear or equipercentile equating. The speed component will influence the direction of the verbal true score dimension and consequently have an impact on conventional equating. For formula scored tests, such as the GRE Aptitude Test, the assumption that examinees will respond only to those items that they have reached is more tenable than it is for number-right scored tests. To the extent that this assumption is reasonable, the convention we chose in estimating parameters with LOGIST, coding all consecutively omitted items at the end of an examinee's answer sheet as not reached, should mitigate the impact of a speed component on the parameter estimates. Hence, a priori we expected a differential effect of speededness on the IRT and conventional equatings of the verbal scale.

In sum, the four factor analytic investigations of the GRE Aptitude Test strongly suggest that separation of verbal items into reading comprehension items and vocabulary (discrete verbal) items would yield two sets of items that are more homogeneous than the single set of all verbal items. The studies also suggest that data interpretation items should be separated from other quantitative items and that technical reading comprehension items may define another distinct set. Doubts about the practical significance of these dimensions, coupled with the fact that there are too few items to permit stable linking of ability scales through IRT difficulty parameters, led us to conclude that separate scales for data interpretation and technical reading comprehension should not be established for the GRE IRT feasibility research.

TEST FORMS AND SAMPLES

Test Forms

Four operational forms of the GRE Aptitude Test were used in this study: ZGR1, K-ZGR2, K-ZGR3 and 3CGR1. The three Z-forms are composed of four separately timed operational sections:

Section	Item Type	Timing in Minutes	Number of Items
I.	Verbal	50	80
	discrete verbal		55
	reading comprehension		25
II.	Quantitative	50	55
	quantitative comparison		30
	data interpretation & regular math		25
III.	Analytical	25	40
	analysis of explanations		40
IV.	Analytical	25	30
	logical diagrams		15
	analytical reasoning		15

The fifth section of each of the three Z-forms contained a 25-minute set of experimental pretest items. A total of seven pretest sections were employed in this study to link the three Z-forms of the GRE Aptitude Test. Table 1 contains pertinent information about these seven pretest forms: their pretest designation, item type, number of items, number of items used for linking. While the first three columns of Table 1 are self-explanatory, the fourth column requires elaboration.

All pretest items are newly written items or revised items that appear in the test in order to develop item statistics for use in assembling operational test forms that have prespecified psychometric characteristics. For the purpose of this study, these experimental sections provided the item parameter links between the three Z-forms under study. For example, the items in pretests B41 and B43 were used to link the verbal ability scales of the GRE Aptitude Test. (Further discussion of linking of IRT ability scales is deferred to the section on linking of ability scales through item difficulty parameters.)

Since these pretest items were being used for the purpose of linking IRT ability scales, which is a prerequisite for IRT score equating of the three Z-forms, it was important to discard items with unacceptable psychometric characteristics. The numbers appearing in the fourth column of Table 1 are the numbers of items that survived the screening procedure for discarding items with unacceptable psychometric characteristics.

The fourth operational form, 3CGR1, is also composed of four separately timed operational sections:

Section	Item Type	Timing in Minutes	Number of Items
I.	Verbal	50	75
	discrete verbal		53
	reading comprehension		22
II.	Quantitative	50	55
	quantitative comparisons		30.
	data interpretation & regular math		25
III.	Analytical	25	36
	analysis of explanations		36
IV.	Analytical	25	30
	logical diagrams		15
	analytical reasoning		15

Form 3CGR1 was administered with six different 25-minute fifth sections. The items in these sections were not experimental pretest items. Instead, they were items taken from the four operational sections of form ZGR1. Table 2 lists the six fifth sections of 3CGR1, the number of items in the section, and the section of ZGR1 from which they were drawn.

In addition to the seven pretest sections listed in Table 1, form ZGR1 was administered with six other section V's at the same administration at which form 3CGR1 was administered with the six section V's listed in Table 2. Table 3 lists these six fifth sections of form ZGR1, indicating the number of items in the section, and the section of 3CGR1 from which they were drawn.

Inspection of Tables 2 and 3 reveals that each operational item from form ZGR1 appears in one of the six section V's of form 3CGR1 and each operational item from 3CGR1 appears in one of the six C-subforms of form ZGR1. This commonality of items was used to study position effects.

Samples

The various forms of the GRE Aptitude Test used in this study were administered at four different times of year. Table 4 identifies the administration date at which each form was administered, and the sample sizes used in this research. Note that form ZGR1 was administered twice: in February 1980 with the B-series of pretests that were shared with forms K-ZGR2 and K-ZGR3, and in June 1980 with the C-series of section V's that contained operational items from form 3CGR1. Form K-ZGR2 was administered in December 1979 to a high ability population containing scientifically oriented candidates competing for National Science Foundation fellowships (although the fellowship candidates made up only about 5 percent of the December examinees, the potential effect of this group was considered important).

Table 1

Experimental Sections for Forms ZGR1, K-ZGR2 and K-ZGR3

Designation	Item Type	Number of Items	Number of Items Used for Linking
B41	Discrete Verbal	55	47
B43	Reading Comprehension	25	20
B46	Quantitative Comparison	40	33
B48	Regular Math	25	23
B50	Data Interpretation	16	12
B52	Analysis of Explanations	50	39
B53	Logical Diagrams and	16	11
	Analytical Reasoning	15	11

Table 2

Six Section V's for Form 3CGR1

Designation	Item Type	Number of Items	Location in ZGR1
C41	Verbal	39	Section I
C42	Verbal	41	Section I
C43	Quantitative	27	Section II
C44	Quantitative	28	Section II
C45	Analytical	40	Section III
C46	Analytical	30	Section IV

Table 3

Six Section V's for Form ZGR1

Designation	Item Type	Number of Items	Location in 3CGR1
C47	Verbal	37	Section I
C48	Verbal	38	Section I
C49	Quantitative	27	Section II
C50	Quantitative	28	Section II
C51	Analytical	36	Section III
C52	Analytical	30	Section IV

Table 4

Description of Samples Used in this Research

Administration Date	Forms	Experimental Section	Sample Size	Formula Score Means and Standard Deviations			
				Experimental		Operational*	
				x	s	x	s
December 1979	K-ZGR2$_{B41}$	V	2315	22.23	9.36	35.93	15.84
	K-ZGR2$_{B43}$	V	2259	5.96	4.02	36.01	15.71
	K-ZGR2$_{B46}$	Q	2333	14.76	7.95	27.40	10.93
	K-ZGR2$_{B48}$	Q	2265	14.17	5.69	26.88	10.83
	K-ZGR2$_{B50}$	Q	2262	7.03	2.85	27.10	10.58
February 1980	ZGR1$_{B41}$	V	2268	20.48	9.46	32.03	15.71
	ZGR1$_{B43}$	V	2207	5.29	3.87	31.86	15.90
	ZGR1$_{B46}$	Q	2274	13.72	7.39	24.63	9.88
	ZGR1$_{B48}$	Q	2216	13.52	5.55	24.84	9.97
	ZGR1$_{B50}$	Q	2231	6.48	2.86	24.55	9.93
April 1980	K-ZGR3$_{B41}$	V	2429	20.32	9.39	33.10	14.61
	K-ZGR3$_{B43}$	V	2406	5.07	3.93	33.08	14.70
	K-ZGR3$_{B46}$	Q	2426	13.12	7.52	25.19	11.16
	K-ZGR3$_{B48}$	Q	2414	13.24	5.75	25.23	11.26
	K-ZGR3$_{B50}$	Q	2414	6.56	2.87	24.93	11.22
June 1980	ZGR1$_{C47}$	V	2483	13.23	8.01	31.61	15.86
	ZGR1$_{C48}$	V	2486	14.62	8.10	31.53	16.30
	ZGR1$_{C49}$	Q	2498	11.94	6.43	24.46	10.47
	ZGR1$_{C50}$	Q	2484	12.88	5.93	24.26	10.34
	ZGR1$_{C51}$	A	2488	18.73	9.13	32.89	15.21
	ZGR1$_{C52}$	A	2482	14.14	6.98	32.69	15.66
	3CGR1$_{C41}$	V	1489	15.54	8.42	30.17	15.38
	3CGR1$_{C42}$	V	1495	15.91	8.80	30.43	15.55
	3CGR1$_{C43}$	Q	1487	11.65	5.59	24.94	11.51
	3CGR1$_{C44}$	Q	1497	12.27	5.43	24.41	11.74
	3CGR1$_{C45}$	A	1526	24.26	11.75	28.86	15.41
	3CGR1$_{C46}$	A	1476	15.92	7.19	28.52	14.87

*Operational-formula raw scores are for the operational section corresponding to the pretest section listed in column three.

PARAMETER ESTIMATION AND ITEM LINKING

Item Calibration Procedures

Data from four administrations were used in this research to assess the feasibility of using item response theory as a psychometric model for the GRE Aptitude Test. A total of 10 different item types (see Table 5) were administered within each form. All item parameter estimates and ability estimates were obtained with the program LOGIST (Wood, Wingersky & Lord, 1978). The function of LOGIST is to estimate, for each item, the three item parameters of the three-parameter logistic model: a_g (discrimination), b_g (difficulty), and c_g (pseudoguessing parameter); and, for each examinee, θ (ability). The following constraints were imposed on the estimation process: a_g was restricted to values between 0.01 and 1.50 inclusive, except for analytical item calibrations where the upper bound was 1.20; the lower limit for θ was -7; and c_g was restricted to values between 0.0 and 0.5. Additionally, each examinee was required to have responded to at least 20 items in order to insure stable θ estimates. Choosing appropriate constraints is a complex procedure, but necessary to speed convergence and produce stable estimates.

For each administration, from four to six different item calibrations were performed. Table 5 shows the relationship between the item types, calibrations and sections of the GRE Aptitude Test. Every item belongs to one item type, but may have been calibrated with more than one set of items (e.g., every analogy item was calibrated with all verbal items and with discrete verbal items only), may have been calibrated more than once with the same set of items in the same relative positions (e.g., all quantitative items on form ZGR1 were calibrated twice, once when administered in February 1980 and once when administered in June 1980), or may have been calibrated with the same set of items in different positions (e.g., every verbal item appearing in section I of form ZGR1 also appeared in a section V of form 3CGR1).

Item Linking Plan

Any meaningful comparisons between item parameters or ability estimates require a common metric (Dorans, 1979). Consequently, the linking plan used to place the item and ability estimates on a common scale is an important aspect of this research. Figures 2 through 5 depict the item linking plans employed. The various verbal item linkings are portrayed in Figures 2 through 4. Figure 2 displays the strategy used to link all the verbal item types. Each of the four test forms is represented by a rectangle attached to a square. The rectangle contains information about the operational section of the test: section number and the number of items. The square contains information about section V (experimental) of the test: subform designation and number of items. Test forms are ordered vertically by administration date. For example, test form ZGR1, administered in February 1980, is represented by the operational rectangle containing I, section number, and 80, number of items, connected to the experimental test square containing B41 and B43, subform designations, and

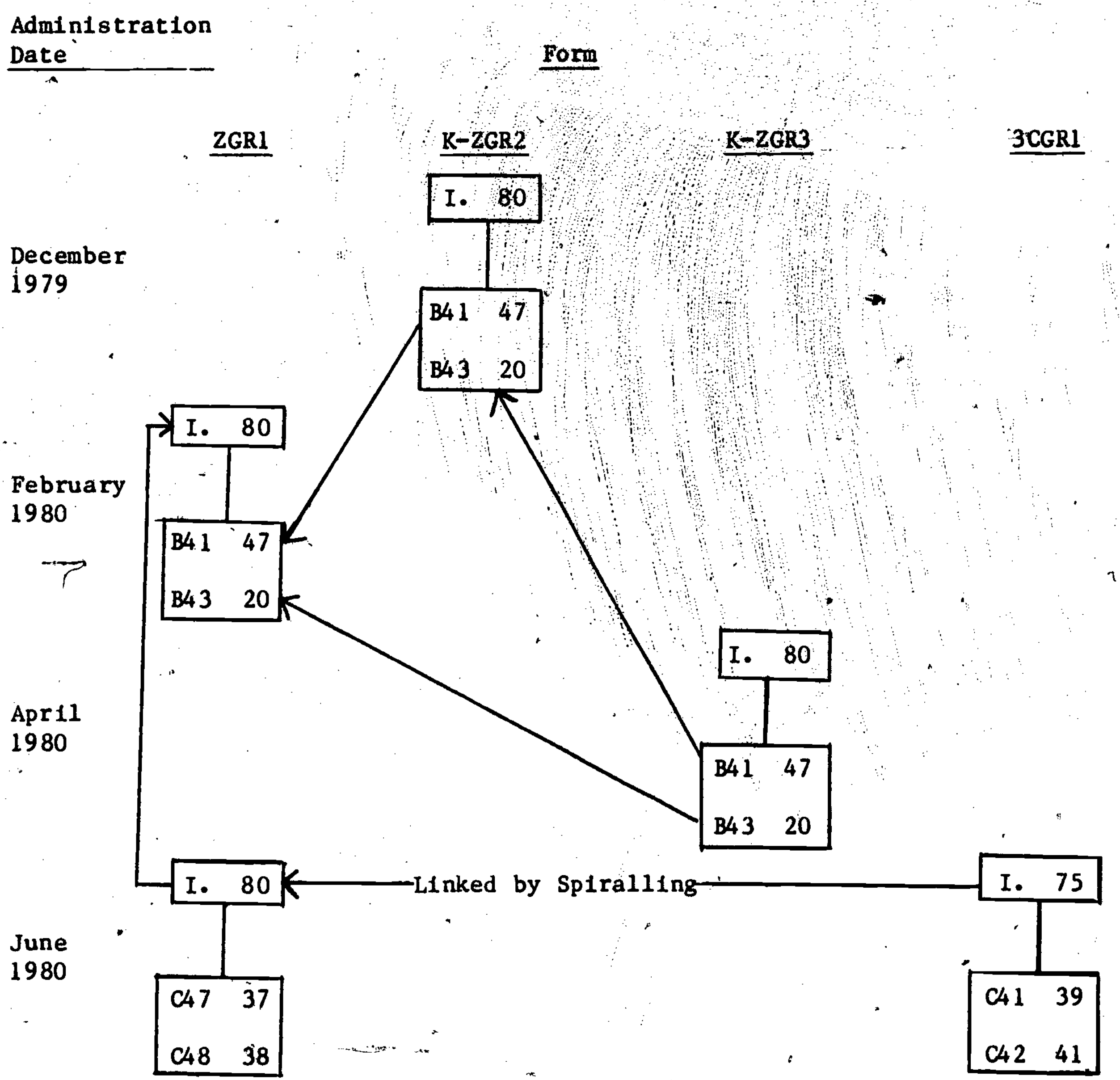
Table 5

Relationships Between Item Type, Calibrations, and Sections of the GRE Aptitude Test

	Calibrations	
Item Types	Subsections	Sections
analogies	discrete, verbal	verbal
antonyms		
sentence completions		
reading comprehension	reading comprehension	
regular mathematics		quantitative
data interpretation		
quantitative comparison	quantitative comparison	
analysis of explanations		analytical
logical diagrams		
analytical reasoning		

Figure 2

IRT Linking Plan for Verbal Scales of GRE Aptitude Test



47 and 20, number of linking items in pretests B41 and B43, respectively. Note in Figure 2 that Form ZGR1, administered in June 1980, is the base form, that forms K-ZGR2 and K-ZGR3 are linked to the February 1980 administration of form ZGR1 through pretest sections B41 and B43, that the February 1980 administration of form ZGR1 is linked to its June 1980 administration by the 80 operational items of section I, and that form 3CGR1 is linked to form ZGR1 through spiralling at the June 1980 administration.

As stated at the end of the factor analytic review, a decision was made to separate the reading comprehension items from the discrete verbal items to establish distinct reading comprehension and discrete verbal scales. Hence, in addition to having been calibrated with all verbal items, each discrete verbal item was calibrated with discrete verbal items only and each reading comprehension item was calibrated with reading comprehension items only. After calibration, each discrete verbal item set was placed on its parent verbal scale. These discrete verbal to verbal scale linkings are depicted in Figure 3.

Each combination (test form/administration date) is represented by two rectangles in Figure 3: an all verbal rectangle and a discrete verbal rectangle. Each all verbal rectangle is partitioned into a three-by-three matrix. The first column of each of these matrices contains a section designation. The second and third columns contain the number of reading comprehension items and the number of discrete verbal items respectively. For example, the matrix for the June 1980 administration of Form ZGR1 indicates that Section I contained 25 reading comprehension items and 55 discrete verbal items, the C47 experimental section contained 11 reading comprehenison items and 26 discrete verbal items, and the C48 experimental section contained 11 reading comprehension items and 27 discrete verbal items.

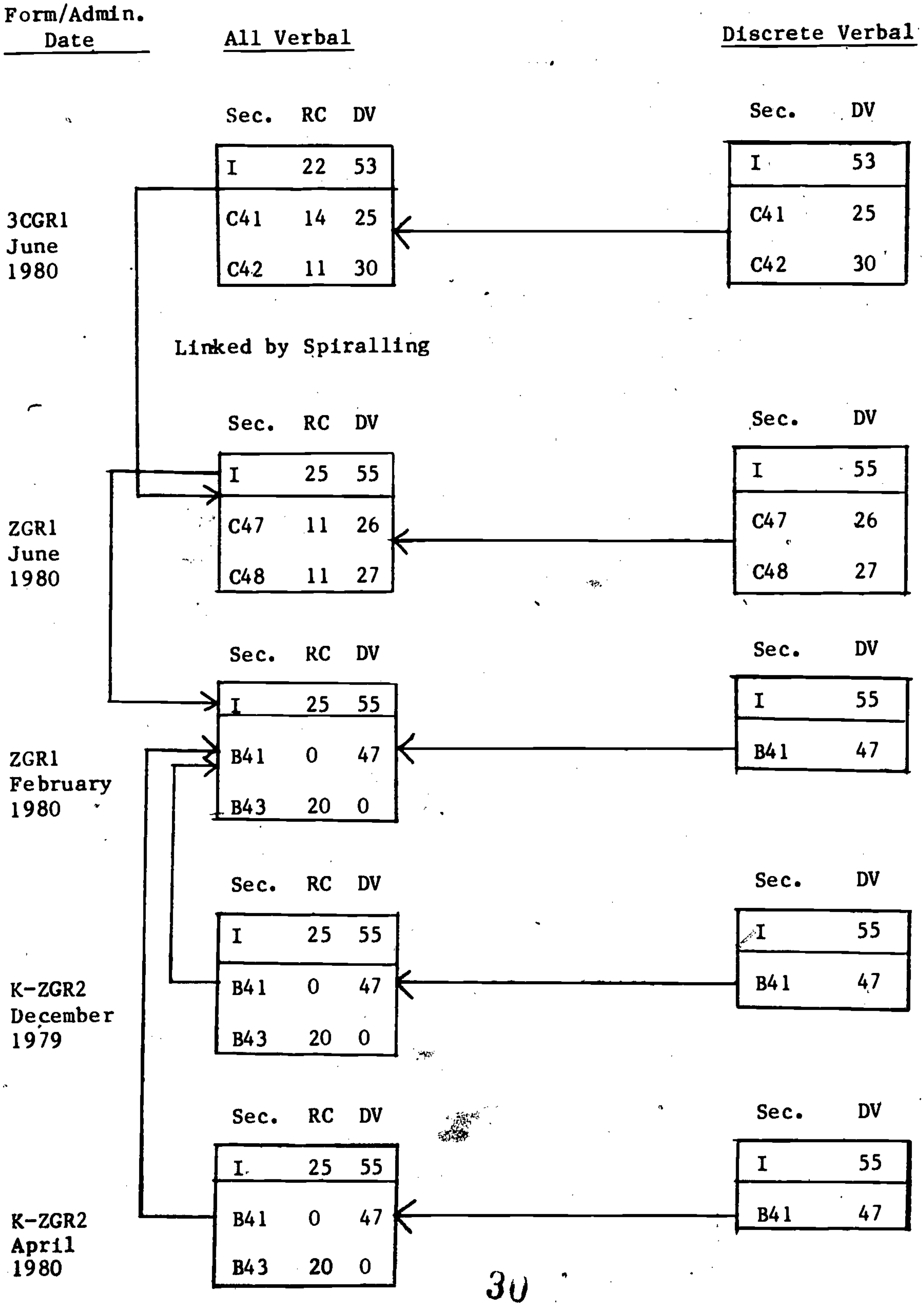
For each all verbal rectangle there is a corresponding discrete verbal rectangle that contains the position of the information contained in the all verbal rectangle that defines the common item link, i.e., the section designation and the number of common discrete verbal items. The arrows in the figure define the direction of the various linkings, which all culminate at the ZGR1 (6/80) all verbal rectangle. For example, the two ZGR1 (2/80) rectangles indicate that the discrete verbal item and ability parameters of ZGR1 (2/80) were placed on the verbal base scale of form ZGR1 (6/80) by a ZGR1 (2/80) to ZGR1 (6/80) all verbal linking via the 80 operational items of section I, and a ZGR1 (6/80) to ZGR1 (6/80) discrete verbal to all verbal linking via the 55 discrete verbal items of section I and the 47 discrete verbal items of pretest B41.

Figure 4 depicts the IRT linking plan for reading comprehension scales of the GRE Aptitude Test. It is similiar in format to Figure 3. Each reading comprehension rectangle contains the section designations and number of reading comprehension items used to place each reading comprehension scale on its parent verbal scale.

Figure 5 depicts the IRT linking plan for the quantitative scales of the GRE Aptitude Test. It is similiar in format to Figure 2.

Figure 3

IRT Linking Plan for Discrete Verbal Scales of GRE Aptitude Test



Form/Admin. Date | All Verbal | Discrete Verbal

3CGR1 June 1980

Sec.	RC	DV
I	22	53
C41	14	25
C42	11	30

Sec.	DV
I	53
C41	25
C42	30

Linked by Spiralling

ZGR1 June 1980

Sec.	RC	DV
I	25	55
C47	11	26
C48	11	27

Sec.	DV
I	55
C47	26
C48	27

ZGR1 February 1980

Sec.	RC	DV
I	25	55
B41	0	47
B43	20	0

Sec.	DV
I	55
B41	47

K-ZGR2 December 1979

Sec.	RC	DV
I	25	55
B41	0	47
B43	20	0

Sec.	DV
I	55
B41	47

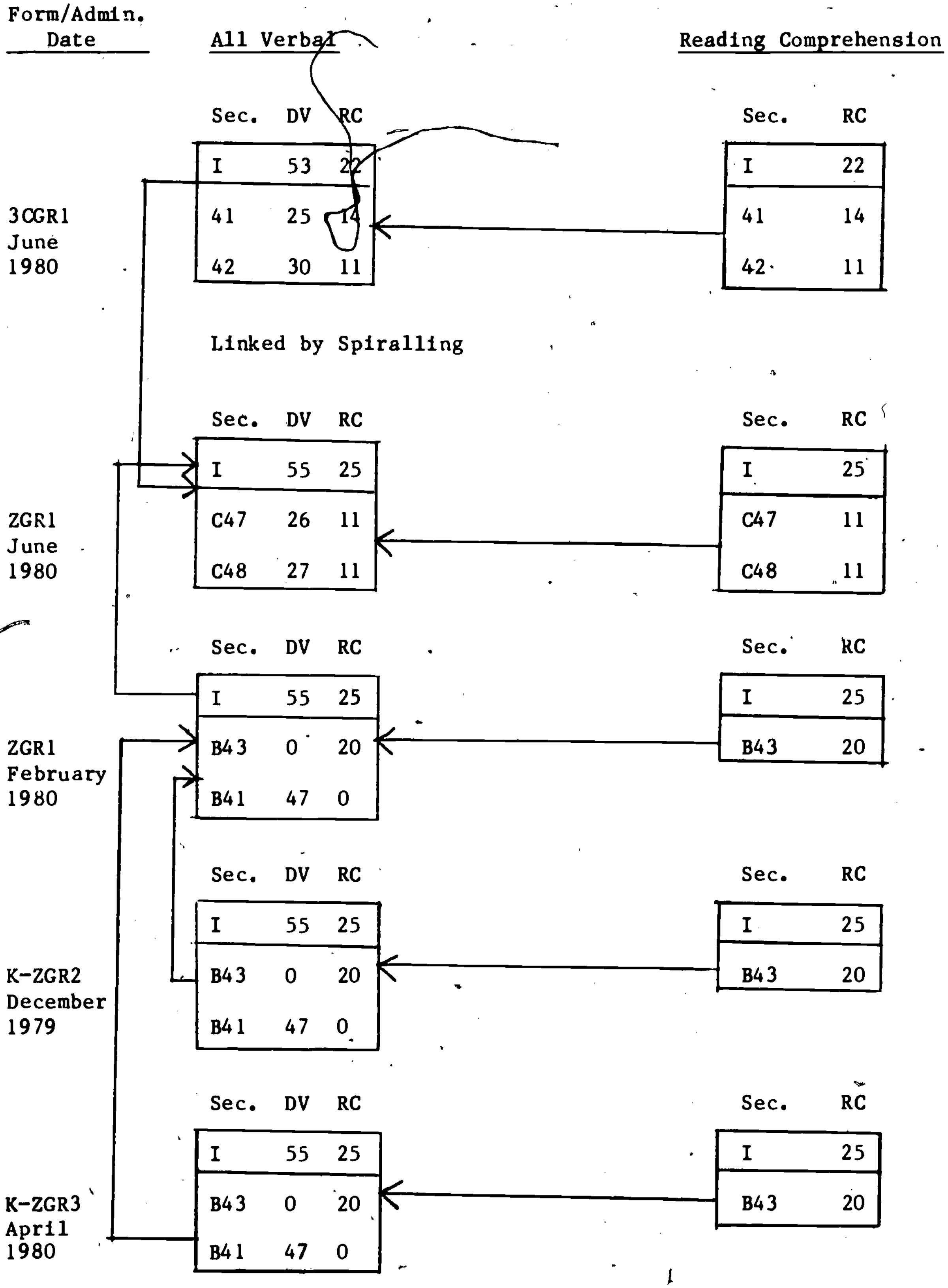
K-ZGR2 April 1980

Sec.	RC	DV
I	25	55
B41	0	47
B43	20	0

Sec.	DV
I	55
B41	47

Figure 4

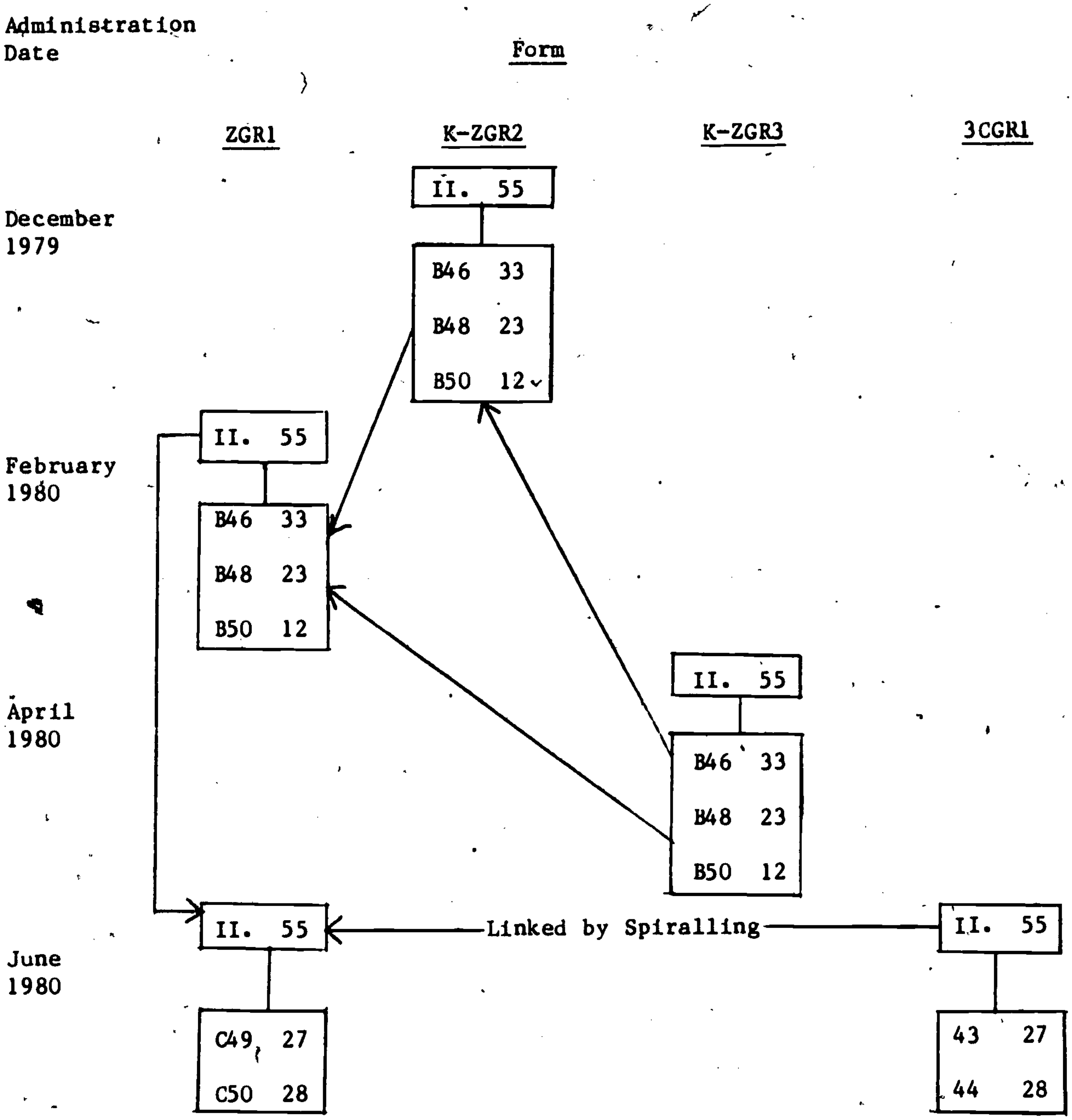
IRT Linking Plan for Reading Comprehension Scales of GRE Aptitude Test



Form/Admin. Date	All Verbal			Reading Comprehension	
	Sec.	DV	RC	Sec.	RC
3CGR1 June 1980	I	53	22	I	22
	41	25	14	41	14
	42	30	11	42	11
ZGR1 June 1980	I	55	25	I	25
	C47	26	11	C47	11
	C48	27	11	C48	11
ZGR1 February 1980	I	55	25	I	25
	B43	0	20	B43	20
	B41	47	0		
K-ZGR2 December 1979	I	55	25	I	25
	B43	0	20	B43	20
	B41	47	0		
K-ZGR3 April 1980	I	55	25	I	25
	B43	0	20	B43	20
	B41	47	0		

Figure 5

IRT Linking Plan for Quantitative Scales of GRE Aptitude Test



IRT Linking Procedures

Two procedures were used to place item parameter and ability estimates on the same metric: spiralling of test forms and a common item linking procedure developed by Lord and Stocking (Petersen, Cook, & Stocking, 1981). Spiralling of test forms at the June 1980 administration of the GRE Aptitude Test was used to link parameter estimates on Form 3CGR1 to parameter estimates on the base form, Form ZGR1. The common item linking procedure was used for all other item linkings.

Linking by spiralling assumes that alternating forms administered to examinees results in a random assignment of forms to examinees. Since large equivalent groups take each form, the distributions of ability in the two groups should be the same, and separate parameterizations based on these two random groups via separate LOGIST runs should produce a single ability metric.

The Lord-Stocking linking procedure produces robust estimates of location and scale of each distribution of item difficulties and an equation based on these robust estimates of location and scale. This equation is used to convert the parameter estimates of a set of items on one form from the arbitrary metric produced by the LOGIST calibration of those items on that form to the base metric resulting from the calibration of the June 1980 administration of Form ZGR1 items. A step-by-step description of the linking of Form K-ZGR3 verbal items is used to illustrate the procedure.

From Figure 2, we see that the K-ZGR3 items are linked to the base form ZGR1, administered in June 1980, via two pathways. The first step in both pathways is to link the February 1980 administration of ZGR1 verbal items to the June 1980 administration of ZGR1 via the 80 shared items from section I. The end result of this procedure is the transformation of parameter estimates from the February 1980 administration of ZGR1 to the base metric of the June 1980 administration of ZGR1. One pathway directly links K-ZGR3 to the transformed ZGR1 (of 2/80) metric via the 67 shared items of pretest sections B41 and B43. The second pathway links K-ZGR3 to ZGR1 through Form K-ZGR2. The first step in both pathways, the linking of the two ZGR1 administrations, will be used to illustrate the Lord-Stocking procedure.

We start with two sets of item difficulty estimates, one from each administration of ZGR1. Each difficulty estimate is weighted by the reciprocal of its squared standard error of estimate; for each item, the larger estimate of its two standard errors of estimate (from the two estimates of item parameters) is used. Then the means and standard deviations of these weighted item difficulty estimates are computed and used to obtain the conversion line that converts the mean and standard deviation of the February 1980 estimates to the mean and standard deviation of the June 1980 estimates. At this point the process becomes iterative. The perpendicular distances of the item difficulty points from this conversion line are computed, and then biweights (Mosteller & Tukey, 1977, p. 205) for these distances are obtained. These biweights are then applied to the reweighted points and a new conversion line is produced. The distance, biweight, reweighting, and new conversion line cycle is repeated until the maximum change in perpendicular distance is less than some criterion. The last conversion line produced by this process is then used to place the February 1980 items on the June 1980 metric. The results of the linking of the two administrations of verbal items appear in Table 6. The final conversion line has a slope of .9960 and an intercept of .0092.

Results of Linking Test Forms

Tables 6, 7, 8, and 9 contain the results of the linkings depicted in Figures 2, 3, 4 and 5, respectively. The verbal linking results are presented in Table 6. Perusal of this table reveals that, with the exception of Form K-ZGR2, the scale transformations produced only slight changes in location and scale; that is, α and β approach 1.0 and 0.0, respectively.

The four weighted correlations in Table 6 are all very high, as should be expected. Visual evidence of this can be seen in Figure 6, which is a scatter plot of difficulties for the 80 common verbal items used to link the two ZGR1 administrations. The noticeable outlier in this plot is item 78, which had a b of -.288 on ZGR1 (6/80) and a transformed difficulty of .729 on ZGR1 (2/80). It should be noted that an outlier as extreme as this gets very little weight compared to the other data points. Except for this peculiar outlier, Figure 6 is typical of difficulty scatter plots for all four verbal linkings.

The review of factor analytic research on the GRE Aptitude Test suggested separation of verbal items into mutually exclusive discrete verbal and reading comprehension sets. Table 7 contains the results of the six discrete verbal linkings, which placed the discrete verbal items onto the metric of the verbal items after the latter had been transformed to the base metric of ZGR1 (6/80). With the exception of the K-ZGR2 transformation, only slight shifts in scale and location were required to convert the discrete verbal scales to the metric of their parent verbal scales.

Table 8 contains the results for the six reading comprehension linkings. A striking feature of Table 8 is the consistent large value for the intercept when scaling reading comprehension items to the verbal scale. (The K-ZGR2 intercept is somewhat larger than the other five.) This finding should not influence model fit or equating and is easily explained. The examinees whose responses were used to estimate the item parameters in the reading comprehension calibrations were more able than those examinees whose responses were used in the verbal calibrations. This is due to our choice of a minimum number of items to which examinees must have responded in order to be included in the calibration procedure.

Consider the reading comprehension calibrations for form K-ZGR2. Examinees who responded to fewer than 20 of the 45 reading comprehension items were dropped from the calibration, i.e., their item responses were

Table 6

Results of Verbal Item Linkings: Correlations (r) Between Weighted Difficulties; and Conversion Equation Parameters, Slope (α) and Intercept (β)

"Old" Form			"New" Form

ZGR1(6/80)	n=80
	α=.9960
		r= .9968
		β= .0092
			ZGR1(2/80)

ZGR1(2/80)T*	n=67
	α=.9401
		r= .9912
		β= .1776
			K-ZGR2(12/79)

ZGR1(2/80)T	n=67
	α=.9906
		r= .9942
		β=-.0282
			K-ZGR3(4/80)

K-ZGR2(12/79)T	n=67
	α=.9907
		r= .9873
		β=-.0338
			K-ZGR3(4/80)

Table 7

Results of Discrete Verbal Item Linkings: Correlations (r) Between Weighted Difficulties; and Conversion Equation Parameters, Slope (α) and Intercept (β)

Form		

ZGR1(6/80)	n=108
	α=1.0005
		r= .9996
		β= .0377

ZGR1(2/80)T	n=102
	α= .9823
		r= .9995
		β= .0582

K-ZGR2(12/79)T	n=102
	α= .9215
		r= .9994
		β= .2406

K-ZGR3(4/80)T1	n=102
	α= .9952
		r= .9997
		β= .0032

K-ZGR3(4/80)T2	n=102
	α= .9953
		r= .9997
		β=-.0025

3CGR1(6/80)	n=108
	α=1.0143
		r= .9996
		β= .0228

*A T suffixed to the "old" form designation indicates the transformation is to scale via an "old" form whose parameter estimates have already been transformed.

Table 8

Results of Reading Comprehension Linkings: Correlations (r) Between Weighted Difficulties; and Conversion Equation Parameters, Slope (α) and Intercept (β)

Form		

ZGR1(6/80)	n=47
	α=.9528
		r=.9965
		β=.1936

ZGR1(2/80)T	n=45
	α=.9792
		r=.9968
		β=.2250

K-ZGR2(12/79)T	n=45
	α=.9753
		r=.9973
		β=.3333

K-ZGR3(4/80)T1	n=45
	α=.9512
		r=.9947
		β=.1726

K-ZGR3(4/80)T2	n=45
	α=.9514
		r=.9947
		β=.1670

3CGR1(6/80)	n=47
	α=.9594
		r=.9960
		β=.1925

Table 9

Results of Quantitative Linkings: Correlations (r) Between Weighted Difficulties; and Conversion Equation Parameters, Slope (α) and Intercept (β)

"Old" Form			"New" Form

ZGR1(6/80)	n=55
	α=.9549
		r=.9980
		β=.03798
			ZGR1(2/80)

ZGR1(2/80)T	n=68
	α=.9799
		r=.9921
		β=.2495
			K-ZGR2(12/79)

ZGR1(2/80)T	n=68
	α=.9690
		r=.9890
		β=.0485
			K-ZGR3(4/80)

K-ZGR2(12/79)T	n=68
	α=.9860
		r=.9921
		β=.0477
			K-ZGR3(4/80)

Figure 6

Parameter Transformations - b
SCR1 (February) to SCR1 (June)
Verbal

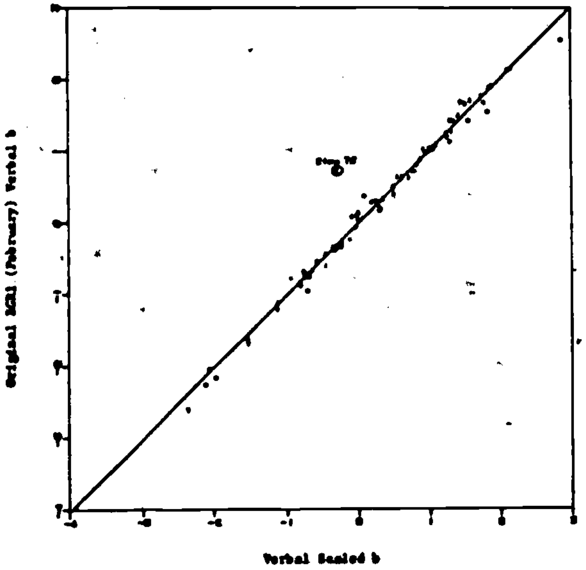


Figure 7

Parameter Transformations - b
K-LCR2 Discrete Verbal to K-LCR2 Verbal

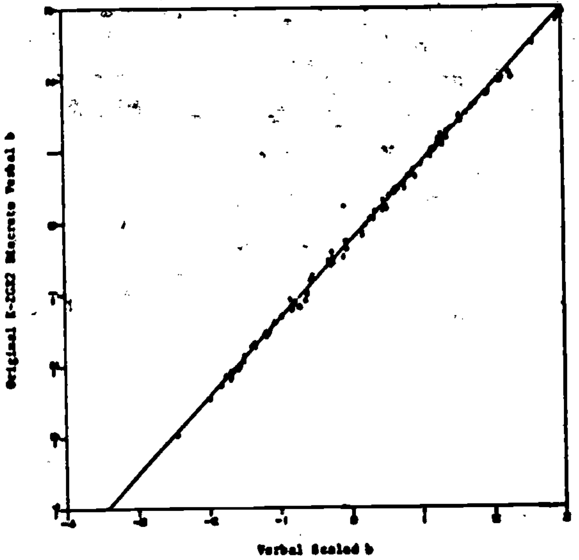


Figure 8

Parameter Transformations - b
K-SCR2 Reading Comprehension to K-SCR2 Verbal

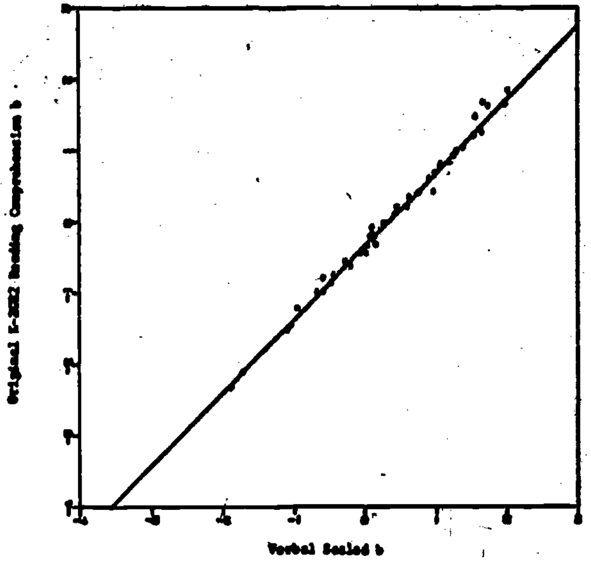
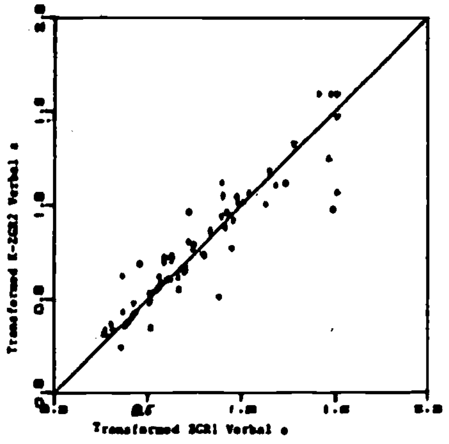


Figure 9

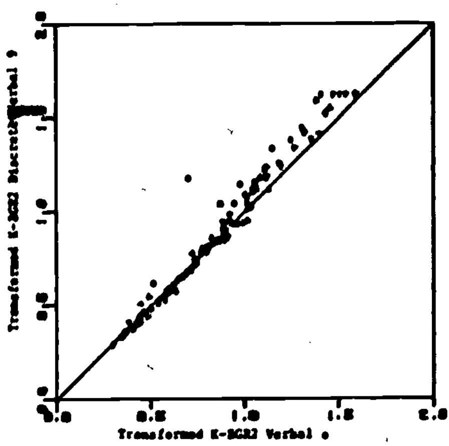
Parameter Transformations -a
K-SCR2 to 2 GR1
Verbal a*



*Since both a estimates have been transformed to scale, the true diagonal indicated in the figure should describe their relationship. Deviation from the diagonal is due to errors of estimation or lack of model fit.

Figure 10

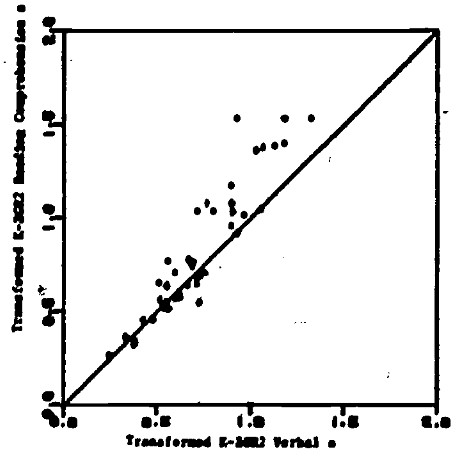
Parameter Transformations -b
K-SCR2 Discrete Verbal to K-SCR2 Verbal *



*Since both a estimates have been transformed to scale, the true diagonal indicated in the figure should describe their relationship. Deviation from the diagonal is due to errors of estimation or lack of model fit.

Figure 11

Parameter Transformations - a
K-SCR2 Reading Comprehension to K-SCR2 Verbal *



*Since both a estimates have been transformed to scale, the true diagonal indicated in the figure should describe their relationship. Deviation from the diagonal is due to errors of estimation or lack of model fit.

ignored and no ability estimates were produced for them. For the verbal calibrations, examinees had to respond to at least 20 of the 145 to 147 items to be retained in the analysis. On the average, approximately 600 more examinees were dropped from the reading comprehension calibrations than were dropped from the verbal calibrations. Since these 600 examinees answered very few items, they were probably mostly examinees of very low ability. Since LOGIST uses an arbitrary ability metric having a mean of zero and a standard deviation of one, the item difficulty estimates obtained in the more able reading comprehension group are lower than the estimates obtained when all verbal items were calibrated. Figure 8, which is a scatterplot of item difficulties for Form K-ZGR2, illustrates this effect. Note that the conversion line for putting reading comprehension item difficulty estimates on the verbal item scale is essentially parallel to the main diagonal. This difference in item difficulty estimates reflects a true difference in ability in the two calibration groups.

Figures 9, 10, and 11 contain typical scatterplots of the transformed-to-scale item discrimination estimates. Figure 9 depicts the relationship between the a's of all the verbal items common to both form K-ZGR2 and form ZGR1. Since each estimated a has been transformed to scale, the points should fall along the true diagonal indicated in Figure 9. Though the scatter is greater than that on the plots of b estimates; there is no evidence of any systematic departure from the diagonal. Figure 10 depicts the relationship between the transformed-to-scale a's from the discrete verbal calibration of form K-ZGR2 with the transformed a's from the all verbal calibration of that form. Figure 11 shows the relationship between the reading comprehension and the all verbal a's. Note the preponderence of points to the left of the main diagonal. The discrimination parameter estimates for the 45 reading comprehension items are higher when calibrated alone than when calibrated with the discrete verbal items, suggesting that two different, though highly correlated, scales are defined by the two different calibrations. Compare this with Figure 10 which shows a much smaller effect for the discrete verbal linkings.

Table 9 contains the results of the quantitative linkings. Examination of this table yields an observation similiar to that produced by examining Table 6: With the exception of form K-ZGR2, the difficulty parameter transformations produced only slight changes in location and scale. The sizeable shift in location for K-ZGR2 may be attributable to the higher ability sample, containing National Science Foundation fellowship candidates, at the December administration of the GRE Aptitude Test. All four correlations in Table 9 are very high.

ASSESSING THE WEAK FORM OF LOCAL INDEPENDENCE: EXAMINATION OF PARTIAL CORRELATIONS AMONG GRE ITEMS CONTROLLING FOR EXAMINEE ABILITY

Implications of Local Independence

The strong form of local independence states that, for a given ability level, item responses are statistically independent. The weak form of local independence states that, for a given ability level, item responses are linearly independent, i.e., uncorrelated. If local independence held and actual ability scores were available, then the partial correlations among items with ability partialled out would be zero. Since the responses to each item go into the ability estimates, however, slightly negative intercorrelations among the items are expected when these ability estimates are partialled out because of part-total contamination (Lord, 1980b).

Theta estimates were read from data sets created by previous LOGIST runs while item responses were read from separate data sets and were recorded as 1 = correct and 0 = incorrect (incorrect responses included, therefore, omitted and not-reached items as well as incorrectly marked items). Biserials and point biserials with either verbal or quantitative ability estimates, and tetrachoric and partial correlations were calculated for items in the verbal and quantitative subtests for two GRE test forms. For each subtest, two runs were made: one with a correction for guessing (Carroll, 1945; Swinton, 1980) and one without this correction. The matrices of partial tetrachoric correlations were then factor analyzed. It was hoped that a linear factor analysis after first removing the variance due to the dominant (and nonlinearly derived) first factor would present a clearer picture than previous factor analytic studies.

Analysis of Partial Correlations

The partial correlations were examined to identify items that correlated highly among themselves (i.e., items that violated the assumption of the weak form of local independence). It was anticipated that an item would be more likely to correlate highly with an item of its own nominal type than it would with other items in the test and that items at the end of a speeded section would be highly intercorrelated. Moreover, it was expected that the percentage of high positive correlations among items of the same type would be greater than the percentage of high correlations for all items in the test. These expectations were borne out, in some cases rather dramatically, and the results will be discussed in the following sections.

The results from the administration of two GRE test forms, ZGR1 (6/80) and K-ZGR2 (12/79), were examined. As previously stated, the latter form was administered to a sample of above average ability, and so some differences in the distributions of correlations were expected. The differences between distributions obtained from the two forms were, however, slight and nonsystematic. Moreover, results from both forms tended to attest to high correlations among technical reading comprehension items on the verbal subtest and among data interpretation items on the quantitative subtest, and some of the less marked results were also similar across test forms.

Correction for guessing. As stated above, the correlations were obtained both with and without a correction for guessing. When a correction for guessing was made, an initial set of chance-level parameters (equal to .20 for the 80 five-choice verbal items and the 25 five-choice quantitative items and equal to .25 for the 30 remaining four-choice quantitative items) was used. These initial estimates were adjusted downward, based on the data, for some items in order to avoid nonsingular correlation matrices. The overall effect of the correction for guessing was to spread out the distribution of partial correlations. It was suspected that in some cases the procedure might have overcorrected for guessing since some partial correlations greater than 1.0 or less than -1.0 were obtained. Both tetrachorics and biserial correlations were corrected for guessing and the result on the partial correlations, which involve ratios containing both tetrachorics and biserials, may have been an overcorrection. Alternatively, these extreme correlations might simply be due to sampling error. In either case, the homogeneity of some nominal item types was more apparent after correction for guessing.

Results for the Verbal Subtest

The 80 GRE verbal items were broken down into the following five nominal item types for the purpose of this analysis:

Item type	Number of items
Sentence completions	17
Analogies	18
Antonyms	20
Reading comprehension	14
Technical reading comprehension	11

The first three item types (which comprise the class of discrete verbal items) occurred both at the beginning and at the end of the verbal test, while the reading comprehension and technical reading comprehension items were found in the middle of the test. The placement of the discrete verbal items introduces, therefore, the nuisance factor of speededness. Unusually high partial correlations were found among the final 15 or so items in a separately timed section, regardless of their nominal item type. Certainly, this is in part a result of the fact that almost half the examinees did not reach the final items and that those who did attempt these items tended to get them correct. The speededness factor, therefore, complicates the analysis, as does the large number of systematic omissions for some reading passages. Both of these factors will be considered as we turn to the results for each nominal item type.

Table 10

Factor Pattern and Intercorrelations Among
Residual Factors Extracted from Form ZGR1
Verbal Item Correlation Matrix in which
Overall Verbal Ability Estimates Have Been Partialled Out

Item Type	Item Position	Factor I	Factor II
Sentence Completion	1	-0.053	-0.023
	2	0.027	-0.032
	3	-0.079	-0.009
	4	0.063	0.025
	5	-0.269	-0.108
	6	-0.018	-0.035
	7	0.035	0.017
	8	-0.031	-0.063
	53	-0.125	-0.194
	54	0.036	0.011
	55	0.114	-0.051
	56	-0.741	-0.231
	57	0.130	-0.013
	58	0.103	0.012
	59	-0.037	-0.026
	60	0.496	-0.054
	61	0.388	0.032
Analogies	9	-0.065	-0.024
	10	0.031	-0.100
	11	0.539	-0.094
	12	-0.183	-0.302
	13	0.286	-0.143
	14	0.093	-0.061
	15	0.309	-0.055
	16	0.194	0.005
	17	0.235	-0.075
	62	0.343	0.012
	63	-0.032	-0.000
	64	-0.158	-0.059
	65	-0.430	-0.137
	66	-0.186	0.054
	67	0.065	0.100
	68	0.031	0.077
	69	0.071	0.050
	70	-0.026	0.034

Table 10 continued

Item Type	Item Position	Factor I	Factor II
Antonyms	18	-0.114	0.079
	19	-0.018	0.078
	20	0.038	0.136
	21	0.011	0.131
	22	0.041	0.152
	23	-0.148	0.188
	24	0.164	0.220
	25	0.045	0.231
	26	0.089	0.202
	27	-0.106	0.699
	71	-0.160	0.837
	72	-0.243	0.909
	73	-0.147	0.692
	74	-0.035	0.776
	75	-0.150	0.834
	76	0.018	0.798
	77	0.012	0.668
	78	0.277	0.001
	79	0.278	0.065
	80	0.420	-0.011
Reading Comprehension	28	0.258	0.030
	29	0.400	-0.010
	30	0.250	0.048
	34	0.393	0.060
	35	0.467	0.010
	36	0.402	0.027
	37	0.437	0.022
	38	0.465	0.042
	39	0.480	0.021
	40	0.531	-0.034
	41	0.580	0.023
	42	0.630	0.024
	43	0.571	-0.046
	44	0.476	0.096
Technical Reading Comprehension	31	0.592	-0.011
	32	0.579	-0.022
	33	0.673	0.008
	45	0.671	-0.081
	46	0.657	-0.022
	47	0.699	-0.088
	48	0.568	-0.017
	49	0.608	-0.066
	50	0.682	-0.090
	51	0.586	-0.057
	52	0.588	-0.046

	Factor I	Factor II
Factor I	1.000	.154
Factor II	.154	1.000

Table 11

Factor Pattern and Intercorrelations Among Residual Factors Extracted from Form K-ZGR2 Verbal Item Correlation Matrix in which Overall Verbal Ability Estimates Have Been Partialled Out

Item Type	Item Position	Factor I	Factor II
Sentence Completion	1	-0.203	-0.041
	2	-0.272	-0.118
	3	-0.185	-0.170
	4	-0.098	0.044
	5	-0.318	-0.099
	6	0.174	-0.034
	7	0.207	-0.037
	8	0.408	-0.007
	53	0.060	-0.105
	54	-0.115	-0.175
	55	-0.153	-0.140
	56	-0.159	-0.215
	57	-0.142	-0.062
	58	0.084	-0.096
	59	-0.005	0.006
	60	0.429	-0.035
	61	0.423	-0.086
Analogies	9	-0.019	-0.080
	10	-0.484	-0.534
	11	-0.148	-0.111
	12	0.325	-0.022
	13	-0.165	-0.230
	14	0.385	0.031
	15	0.275	0.016
	16	0.058	-0.058
	17	0.359	-0.133
	62	0.436	-0.062
	63	0.037	0.039
	64	-0.295	-0.107
	65	-0.277	-0.044
	66	-0.041	0.386
	67	-0.107	0.420
	68	-0.062	0.527
	69	-0.085	0.680
	70	-0.141	0.643
Antonyms	18	-0.207	0.737
	19	-0.163	0.701
	20	-0.065	0.507
	21	0.133	0.251
	22	-0.015	0.069
	23	0.015	0.234

Table 11 continued

Item Type	Item Position	Factor I	Factor II
Antonyms	24	-0.096	0.213
	25	-0.170	0.124
	26	0.004	0.138
	27	0.046	0.131
	71	0.025	0.132
	72	-0.071	0.109
	73	0.098	0.106
	74	0.034	0.262
	75	0.034	0.028
	76	0.048	0.036
	77	0.008	0.023
	78	0.299	0.010
	79	0.196	-0.014
	80	0.283	-0.066
Reading Comprehension	28	0.214	-0.086
	29	0.215	0.007
	30	0.487	0.018
	39	0.270	-0.044
	40	0.424	0.048
	41	0.235	-0.020
	42	0.230	-0.085
	43	0.496	-0.055
	44	0.454	-0.080
	45	0.571	-0.047
	46	0.282	0.119
	50	0.438	-0.020
	51	0.748	-0.019
	52	0.281	0.001
Technical Reading Comprehension	31	0.559	0.094
	32	0.578	-0.043
	33	0.587	0.019
	34	0.625	-0.000
	35	0.611	0.020
	36	0.747	0.099
	37	0.627	0.027
	38	0.585	-0.041
	47	0.515	0.044
	48	0.666	0.001
	49	0.527	0.002

	Factor I	Factor II
Factor I	1.000	.066
Factor II	.066	1.000

Factor analysis of partial correlations. In an effort to summarize the results of the verbal item partial correlation analyses, the partial correlations, not corrected for guessing, were subjected to factor analysis. (The choice of the uncorrected-for-guessing partials was based on the difficulty of estimating communalities using the corrected partials, as well as concern about overcorrection.) Principal factor analysis (Harman, 1976, Chapter 6.3) was used to identify and extract the primary factors of these verbal partial correlation matrices. Since the dominant (nonlinearly derived) ability factor had been partialled out, these remaining factors can be viewed as residual factors that might be systematic sources of local independence violations. Following extraction, these residual factors were rotated to an oblique solution using direct oblimin with Kaiser normalization (Harman, 1976; Chapter 14.4).

The factor pattern (regression weights for predicting common portions of item variables from underlying factors) and factor intercorrelations, following a direct oblimin rotation of a two-factor solution for the Form ZGR1 (6/80) verbal item intercorrelations with overall verbal ability partialled out, appear in Table 10. Clearly, the first factor is defined by the reading comprehension items, primarily the technical reading comprehension items. The second factor appears to be a speed factor as the antonym items appearing at the end of the verbal section mark this factor.

The corresponding results (factor pattern and intercorrelations) for form K-ZGR2 (12/79) appear in Table 11. Again, a two-factor solution was obtained, although the plot of eigenvalues suggested that a one-factor solution might have been sufficient. The first factor was clearly a reading comprehension factor marked by very high loadings for technical reading comprehension items in particular. The definition of the second factor is difficult. It appears to be a mixture of analogy and antonyms, but may well be a composite of noise components, i.e., there may be only one meaningful residual factor, that marked by reading comprehension items. The relative high ability of the group that took Form K-ZGR2 may have caused the speed factor noted in the ZGR1 analysis to dissipate.

In sum, the factor analysis of partial correlation matrices with overall verbal ability partialled out produced results consistent with the visual analysis of partial correlation distributions: evidence for both a technical reading comprehension factor and a nuisance speed factor.

Results for the Quantitative Subtest

The 55 quantitative items were broken down into the following three nominal item types:

Item type	Number of items
Quantitative comparison	30
Regular mathematics	15
Data interpretation	10

The four-choice quantitative comparison items all appear at the beginning of the quantitative section, while regular mathematics and data interpretation items were interspersed in the latter part of the section. It was expected that speededness would prove to be less of a factor for quantitative items than it was for verbal items since, in both test forms, at least 80 percent of the examinees reached item 50 out of 55 items.

Factor analysis of partial correlations. The quantitative partial correlation analyses were summarized by factor analyzing the partial correlations, not corrected for guessing, using principal factor analysis. Factors remaining after the nonlinearly derived dominant quantitative factor had been partialled out can be viewed as residual factors that might be systematic sources of local independence violations. Following extraction, these residual factors were rotated to an oblique solution using direct oblimin with Kaiser normalization.

The factor pattern (regression weights for predicting common portions of the quantitative item variables from underlying factors) and factor intercorrelations, following a direct oblimin rotation of a two factor solution for form ZGR1 (6/80) quantitative item intercorrelations with overall quantitative ability partialled out, appear in Table 12. Both factors are marked by data interpretation items predominantly, suggesting that the two residual factors are different types of data interpretation factors. The corresponding results for form K-ZGR2 (12/79) appear in Table 13. Although two factors were extracted, a single-factor solution was probably sufficient. This first factor is clearly marked by the data interpretation items, while interpretation of the second factor is difficult since it is probably a composite of noise components.

Summary and Synthesis

Principal findings for the verbal subtest. The analysis of partial correlations and the subsequent factor analysis for the verbal subtest uncovered two systematic sources of local independence violation. The reading comprehension items, particularly those pertaining to technical reading passages, retained positive intercorrelations even after overall verbal ability estimates were partialled out. Whether this reading comprehension residual factor is a special skill or simply a function of the fact that sets of items refer to a common passage cannot be absolutely ascertained. Most likely, several influences are at work. In any case, the end result is a violation of local independence.

The second systematic source, most evident in the analysis of form ZGR1, is speededness. Test speededness tends to enhance the partial correlations between items at the end of the test, probably because a self-selected group of higher ability examinees attempt them while those who do not reach them are of lower ability. This ability to perform well on speeded tests is probably related imperfectly to overall verbal ability. In other words, after overall verbal ability has been partialled out,

Table 12

Factor Pattern and Intercorrelations Among
Residual Factors Extracted from Form ZGR1
Quantitative Item Correlation Matrix in which
Overall Quantitative Ability Estimates Have Been Partialled Out

Item Type	Item Position	Factor I	Factor II
Quantitative Comparisons	1	-0.643	-0.094
	2	-0.382	-0.180
	3	-0.318	-0.077
	4	-0.110	-0.176
	5	-0.190	-0.050
	6	-0.216	-0.022
	7	-0.216	-0.039
	8	0.038	-0.124
	9	-0.222	-0.066
	10	-0.023	-0.141
	11	0.121	-0.048
	12	-0.091	-0.058
	13	0.073	-0.150
	14	-0.064	-0.096
	15	0.261	-0.050
	16	-0.045	-0.031
	17	0.308	0.035
	18	0.023	-0.162
	19	0.224	0.022
	20	0.370	-0.154
	21	0.146	-0.056
	22	0.325	-0.051
	23	0.400	-0.133
	24	0.549	-0.008
	25	0.466	-0.286
	26	0.270	-0.009
	27	0.242	0.006
	28	0.228	-0.057
	29	0.147	-0.024
	30	0.122	0.036
Regular Mathematics	31	-0.241	-0.040
	32	-0.108	0.184
	33	-0.116	-0.128
	34	0.048	-0.187
	35	0.317	0.023
	40	-0.189	0.250
	41	-0.199	0.290
	42	0.083	0.158
	43	-0.103	0.270
	44	-0.026	0.202
	51	0.114	0.066

Table 12 continued

Item Type	Item Position	Factor I	Factor II
Regular Mathematics	52	0.526	-0.075
	53	0.514	-0.059
	54	0.042	0.155
	55	-0.005	0.761
Data Interpretation	36	0.369	0.567
	37	-0.151	0.905
	38	0.209	0.462
	39	0.040	0.623
	45	0.156	0.668
	46	0.254	0.411
	47	0.834	0.104
	48	0.406	0.194
	49	0.618	0.166
	50	0.228	0.167

	Factor I	Factor II
Factor I	1.000	.059
Factor II	.059	1.000

Table 13

Factor Pattern and Intercorrelations Among
Residual Factors Extracted from Form K-ZGR2
Quantitative Item Correlation Matrix in which
Overall Quantitative Ability Estimates Have Been Partialled Out

Item Type	Item Position	Factor I	Factor II
Quantitative Comparisons	1	-0.058	-0.122
	2	-0.135	-0.324
	3	-0.084	-0.069
	4	-0.114	-0.391
	5	-0.017	-0.022
	6	-0.102	-0.088
	7	-0.103	-0.177
	8	-0.176	0.127
	9	-0.145	-0.084
	10	-0.060	0.001
	11	-0.047	-0.126
	12	-0.106	-0.064
	13	-0.099	0.157
	14	-0.082	0.132
	15	-0.095	0.059
	16	0.063	0.094
	17	-0.009	0.148
	18	-0.011	0.264
	19	-0.025	0.194
	20	-0.081	0.123
	21	-0.085	0.360
	22	0.038	0.188
	23	-0.027	0.112
	24	0.024	0.110
	25	-0.118	0.003
	26	-0.041	0.138
	27	0.099	0.315
	28	0.005	0.277
	29	0,029	0.176
	30	0.006	0.405
Regular Mathematics	31	-0.022	-0.142
	32	-0.142	-0.385
	33	0.142	-0.385
	34	0.095	-0.040
	35	-0.050	-0.331
	42	-0.068	-0.179
	43	0.181	-0.263
	44	-0.008	0.050
	45	0.121	0.287
	46	0.086	0.288
	51	0.167	0.212

Table 13 continued

Item Type	Item Position	Factor I	Factor II
Regular Mathematics	52	0.020	-0.057
	53	0.218	0.054
	54	0.200	0.333
	55	0.291	0.077
Data Interpretation	36	0.178	-0.062
	37	0.667	-0.241
	38	0.804	-0.214
	39	0.811	-0.398
	40	0.487	0.091
	41	0.447	0.129
	47	0.455	0.274
	48	0.381	0.161
	49	0.335	0.220
	50	0.314	0.192

	Factor I	Factor II
Factor I	1.000	.129
Factor II	.129	1.000

a residual speededness factor remains that systematically influences performance on items appearing at the end of the test.

Principal findings for the quantitative subtest. The analysis of partial correlations and the subsequent factor analyses for the quantitative subtest uncovered a single major source of local independence violations: a factor influencing performance on data interpretation items. On form ZGR1 (6/80), this source seemed to be composed of two components that might be related to differences in data interpretation passages. On form K-ZGR2 (12/79), however, this separation into two components was not evident. In any case, the data interpretation items exhibited positive intercorrelations after general quantitative ability was partialled out. Whatever accounted for these positive correlations is a source of local independence violations.

Synthesis with previous factor analytic results. The partial correlation analyses produced findings consistent with expectations, based on the factor analytic review described in Chapter 2. The earlier factor analytic studies provided strong evidence for the existence of three large global factors in GRE Aptitude Test data: general quantitative ability, vocabulary or discrete verbal ability, and reading comprehension or general verbal reasoning ability. In addition, they provided evidence for the existence of some smaller factors: technical reading comprehension, data interpretation, and verbal speededness factors. The partial correlation analysis just described produced evidence confirming some of these results, most notably results that would suggest violations of local independence.

ANALYSIS OF ITEM-ABILITY REGRESSIONS

Frequently, researchers will try to assess the fit of a latent trait model to real data using a chi-square test or other similar approaches (Wright, 1977). Unfortunately, such tests require expected values that are available only when we know the values of item or people parameters; in the real world we only have estimates of these parameters. These estimates are likely to behave differently from true parameters in a statistical test and would probably increase the probability of a type II statistical error; that is, we would not reject the null hypothesis that the model fits as frequently as we should.

To avoid this problem, a graphical technique and some quantitative summaries of that technique were used in a roughly normative manner to assess the fit of the three-parameter logistic model. This exploratory technique, which will be referred to as analysis of item-ability regressions, compares the regression of the observed proportion of people getting an item correct on estimated θ (empirical regression) with the item response function based on the estimated item parameters (estimated regression) (Hambleton, 1980; Stocking, 1980).

The untransformed ability scale (θ estimated on the metric for which the trimmed calibration sample, examinees with estimated θ between -3.0 and 3.0, has a mean of 0 and a standard deviation of 1) is split into 15 intervals of width .4 in the range -3.0 to +3.0. P_i, the proportion of people in interval i getting the item correct, adjusted for omits, is computed for each in interval. That is,

(3) $$P_i = \frac{n_i^+ + n_i^o/A}{n_i} \text{ , where}$$

n_i^+ is the number of examinees in the i-th interval who got the item correct,

n_i^o is the number of examinees in the i-th interval who omitted the item,

A is the number of alternatives per item,

n_i is the number of examinees in interval i who answered the item or any item subsequent to that item.

The 15 P_i are plotted as squares whose areas are proportional to n_i. For each interval, a line of length $4\sqrt{(PQ/n_i)}$ is plotted, where P and Q are computed from the estimated item response function. The line is centered on the estimated response function. Although this line is a rough estimate of the .95 confidence interval around the item response function, it is not being used as a statistical test. The reasons why this line does not represent the .95 confidence interval include: the use of 2 instead of 1.96 as a coefficient; the use of the inappropriate symmetric normal approximation to the binomial confidence interval around the response function (particularly a problem for extreme values of P); and the use of an interval based on estimated item parameters.

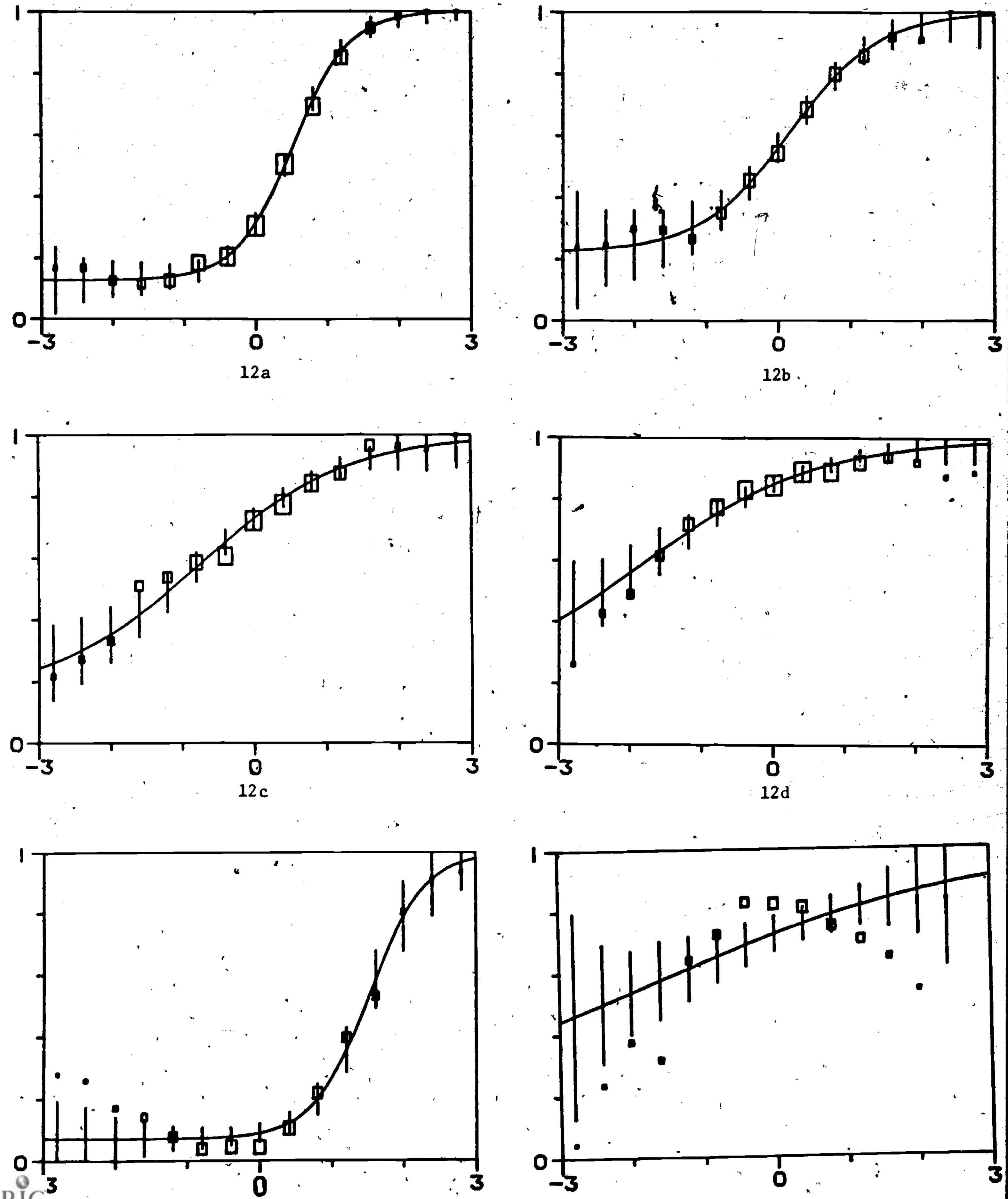
Figures 12a through 12f show six examples of item-ability regressions. The vertical scale in each is the probability of a correct response and ranges from 0 to 1. The horizontal scale is the ability metric and ranges from -3.0 to +3.0. Various attributes of these item-ability regressions relate to model fit. After looking at more than 1,000 of these plots, we decided that a useful summary statistic would be the number of times the proportion of the examinees in an interval responding correctly to the item fell outside the $\pm 2\sqrt{PQ/n}$ interval centered on the response function: that is, the number of times the midpoints of the boxes fell off the vertical lines. Thus, the item-ability regressions in 12a and 12b would each be scored 0, those in 12c and 12d would be scored 2 and 3 respectively, and those in 12e and 12f would be scored 5 and 9.

This analysis is based on 395 verbal, 275 quantitative, and 136 analytical items. The verbal and quantitative items consist of all such operational items from four administrations of the four GRE Aptitude Test forms studied in this research. The analytical items consist of all operational items from forms 3CGR1 and ZGR1.

Table 14 presents cumulative distributions of item scores on the model fit statistic described above. Data are presented for the three major item classifications and their constituent item types. All data presented in this table are based on verbal, quantitative, or analytical calibrations.

To aid interpretation of these data, frequencies of model fit score were collapsed into two categories (1, 2+), and compared across item types with a chi-square test of independence. Table 15 presents these results for the three major item classifications.

Figure 12
Examples of Item Ability Regressions



12a

12b

12c

12d

12e

12f

Table 14

Assessment of Model Fit

		Cumulative Proportion of Items with Model Fit Score Less Than or Equal to:								
Item Type	Number of Items	0	1	2	3	4	5	6	7	8
All Verbal	395	.63	.87	.96	.99	.99+	1.00			
Analogies	90	.62	.84	.93	.98	1.00				
Antonyms	102	.67	.91	.97	.99	.99	1.00			
Sentence Completions	81	.56	.88	.95	1.00					
Reading Comprehension	122	.66	.86	.97	.99	1.00				
All Quantitative	275	.45	.69	.82	.89	.94	.96	.98	.99	1.00
Regular Mathematics	75	.45	.80	.91	.95	.96	.96	.97	.97	1.00
Data Interpretation	55	.56	.80	.90	.94	.98	.98	.98	.98	1.00
Quantitative Comparison	150	.41	.60	.75	.85	.91	.96	.99	.99	1.00
All Analytical	136	.59	.82	.95	.98	.99	.99	.99	.99	1.00
Analysis of Explanations	76	.54	.76	.93	.96	.97	.97	.97	.99	1.00
Logical Diagrams	30	.70	.97	.97	1.00					
Analytical Reasoning	30	.60	.83	.97	1.00					
All Items	806	.56	.80	.91	.96	.98	.99	.99	.99	1.00

Table 15

Comparison of Model Fit for Three Major Item Classifications

	Model Fit Score			
Item Classification	0-1	2+	Total	
Verbal	345	50	395	$\chi^2 = 34.55$
Quantitative	190	85	275	df = 2
Analytical	112	24	136	$p \leq .0001$
Total	647	159	806	

The high χ^2 for Table 15 shows a relationship between broad item classification and model fit. Whether or not the three-parameter logistic model fits any of the item types in an absolute sense, Table 15 shows that some fit more closely than others. In particular, the order of fit seems to be (from best to worst) verbal, analytical, quantitative. Since these differences might be due to specific item types, each broad classification was separately analyzed by specific item type. Table 16 presents these results for verbal items, Table 17 for quantitative items, and Table 18 for analytical items.

Table 16

Comparison of Model Fit for Verbal Item Types

	Model Fit Score			
Item Classification	0-1	2+	Total	
Analogies	76	14	90	$\chi^2 = 2.23$
Antonyms	93	9	102	df = 3
Sentence Completion	71	10	81	$p \leq .5267$
Reading Comprehension	105	17	122	
Total	345	50	395	

Table 17

Comparison of Model Fit for Quantitative Item Types

	Model Fit Score			
Item Classification	0-1	2+	Total	
Regular Mathematics	60	15	75	$\chi^2 = 12.77$
Data Interpretation	40	10	50	df = 2
Quantitative Comparison	90	60	150	$p \leq .0017$
Total	190	85	275	

Table 18

Comparison of Model Fit for Analytical Item Types

	Model Fit Score			
Item Classification	0-1	2+	Total	
Analysis of Explanations	58	18	76	$\chi^2 = 6.16$
Logical Diagrams	29	1	30	df = 2
Analytical Reasoning	25	5	30	$p \leq .0461$
Total	112	24	136	

The four verbal item types presented in Table 16 show no significant difference in model fit. Of the three quantitative item types presented in Table 17, the model fits the quantitative comparison items least well.

One feature of quantitative comparison items is that they all share the same response options and instructions:

> Directions: Each question in this part consists of two quantities, one in Column A and one in Column B. You are to compare the two quantities and on the answer sheet blacken space
> A if the quantity in Column A is the greater;
> B if the quantity in Column B is the greater;
> C if the two quantities are equal;
> D if the relationship cannot be determined
> from the information given.

This might lead to multidimensionality due to the particular correct response of the item. To investigate this, a chi-square test of independence between the keyed response and model fit score (collapsed into two categories) was performed. Results are presented in Table 18. There is no evidence for any response option factors.

Table 19

Comparison of Model Fit for Different
Keyed Responses of Quantitative Comparisons Items

	Model Fit Score		
Keyed Response	0-1	2+	Total
A	23	15	38
B	21	19	40
C	27	12	39
D	19	14	33
Total	90	60	150

$\chi^2 = 2.41$

df = 3

$p \leq .4923$

Alternatively, it could be argued that another type of multidimensionality caused the model fit problem. Perhaps quantitative comparison items themselves are unidimensional, but are tapping a different dimension from the rest of the quantitative items. Factor analytic results, reviewed earlier in this report, did not indicate this to be the case, but the past factor analytic studies used linear models, and item response theory is based on a nonlinear model. A separate quantitative comparison factor could not be ruled out.

To further investigate this, the quantitative comparison items for one form (K-ZGR3) were separately calibrated. Item-ability regressions for items in this calibration could not be affected by multidimensionality inherent across the three quantitative item types. Table 20 compares the model fit for the 30 quantitative comparison items calibrated with the entire quantitative section with that for the items calibrated as a homogeneous subset.

Table 20

Comparison of Model Fit for Homogeneous and Heterogeneous Calibrations of Quantitative Comparison Items

	Model Fit Score		
Calibration	0-1	2+	Total
Quantitative Comparison Only	18	12	30
All Quantitative Items	19	11	30
Total	37	23	60

Since different calibrations of identical items are represented in the two rows of Table 20, a test of independence was not performed. Nonetheless, it seems obvious that any multidimensionality occurs within the item type and not across the three quantitative item types.

Further examination of the items and their directions leads us to hypothesize another type of dimensionality problem. Due to a problem solving response set, some examinees who did not know the answer to a quantitative comparison item might be more likely to answer D, "the relationship cannot be determined from the information given," than others of equal quantitative ability, in which case the poor model fit of these items might be explained. This problem solving response set would contribute to a lack of model fit, regardless of the keyed response. If the correct answer were A, B, or C, some examinees with a given ability would be less likely to pick the correct answer than others because of their propensity for response D. If D were the correct answer, these same examinees would be more likely to pick the correct answer than the model predicted.

Table 18 indicates that the three-parameter logistic model fits analysis of explanations items less well than the other analytical item types. Like quantitative comparisons items, these items all share a single response format:

Directions: For each set of questions, a fact situation and a result are presented. Several numbered statements follow the result. Each statement is to be evaluated in relation to the fact situation and result.

Consider each statement separately from the other statements. For each one, examine the following sequence of decisions, in the order A,B,C,D,E. Each decision results in selecting or eliminating a choice. The first choice that cannot be eliminated is the correct answer.

A Is the statement inconsistent with, or contradictory to, something in the fact situation, the result, or both together? If so, choose A.
If not,

B Does the statement present a possible adequate explanation of the result? If so, choose B.
If not,

C Does the statement have to be true if the fact situation and result are as stated?
If so, the statement is deducible from something in the fact situation, the result, or both together; choose C.
If not,

D Does the statement either support or weaken a possible explanation of the result?
If so, the statement is relevant to an explanation; choose D.

E If not, the statement is irrelevant to an explanation of the result; choose E.

Table 21 presents a test of independence between keyed response and model fit.

Table 21

Comparison of Model Fit for Different Keyed Responses of Analysis of Explanations Items

	Model Fit Score			
Keyed Response	0-1	2+	Total	
A	10	1	11	$\chi^2 = 25.07$
B	7	10	17	df = 4
C	18	1	19	$p \leq .0001$
D	16	0	16	
E	7	6	13	
Total	58	18	76	

Analysis of explanations items keyed B or E were not fit well by the model. In fact, some of the B-keyed items are not monotonically increasing; more able students frequently choose the D response. Figure 12f presents the most extreme example of such an item we have found. Factor analysis (Swinton & Powers, 1980) has provided additional evidence of keyed response specific factors for analysis of explanations items.

In summary, the three-parameter logistic model seems to fit all of the verbal item types and two of the analytical item types, logical diagrams and analytical reasoning, better than the three quantitative item types and the analysis of explanations items. Of the latter four item types, regular mathematics and data interpretation items seem to be fit almost as well as some of the "good fitting" item types. Analysis of explanations items keyed other than B or E were fit by the model quite well (less than 5 percent of the items keyed A, C, or D have a model fit score of 2 or greater), but those keyed B or E have the highest proportion of model fit scores of 2 or greater of any of the item classifications we considered (53%). Quantitative comparison items were the most difficult item type for the three-parameter logistic model to fit.

COMPARABILITY, SENSITIVITY, AND STABILITY OF PARAMETER ESTIMATES

Temporal Stability of Item Parameter Estimates

The operational sections of form ZGR1 were administered twice, once in February and once in June 1980, which allows us to assess the temporal stability of item parameter estimates. Theoretically, the item response function for each item should not be affected by when the item was administered, provided that a common metric has been established. The section on parameter estimation and item linking describes the procedure used to place all item parameter estimates on the same scale. Thus, any discrepancies in item parameter estimates should be due to lack of fit of the three-parameter logistic model because of population shifts or because of errors of estimation. (Though item response theory provides sample invariant parameter estimation, this sample invariance applies to samples (of the same or different ability) from a single population. Population shifts can cause a change in dimensionality.) In this section, the two sets of item parameter estimates (after transformation to a common metric) for form ZGR1 are compared for the verbal calibrations, the discrete verbal calibrations, the reading comprehenison calibrations, and the quantitative calibrations. Tables 22 through 24 summarize these comparisons.

In Table 22, means, standard deviations, and correlations between parameter estimates obtained at both administrations are presented for all 55 discrete verbal items in Section I of form ZGR1. The upper half of the table contains results for the verbal calibrations of these items; the results for the discrete verbal calibrations of these items are presented in the lower half of this table. The parameters a_g, b_g, and c_g are the item discrimination, item difficulty, and pseudoguessing parameters of the three-parameter logistic model. The p_g is an estimate of conventional item difficulty, the proportion of examinees giving a correct response to the item, that is based on the item response function and the marginal distribution of ability for the group of examinees given that item. The p_g can be viewed as a nonlinear bounding transformation of b_g. This bounding transformation was performed for two reasons. First, extreme values of b_g have large standard errors, while extreme values of p_g do not. Second, the Pearson product-moment correlation, used in this section, is sensitive to outliers, and a bounded item difficulty parameter, such as p_g, is less likely to produce troublesome outliers. The p_g values, however, are sensitive to any large differences in group ability and could produce a nonlinear relationship between the p_g estimates of the form ZGR1 items based on the two administrations. As it turned out, the abilities of the two groups were similar enough that nonlinearity was not a problem.

The means and standard deviations to the right of each rectangle are the means and standard deviations of the three item parameters and p_g for the June 1980 administration of form ZGR1, while the summary statistics for the February 1980 administration of form ZGR1 appear under each rectangle. The elements inside the rectangle are correlations between the estimates obtained at the two administrations of form ZGR1. Note that both item difficulty estimates, b_g and p_g, were virtually insensitive

TABLE 22

Correlations and Summary Statistics for Item Parameters and Estimated Proportion Correct for the 55 Discrete Verbal Items of Section I of Form ZGR1

ALL VERBAL CALIBRATION

ZGR1 (2/80)

		a_g	b_g	c_g	p_g	Mean	S.D.
ZGR1 (6/80)	a_g	.914				.899	.312
	b_g		.988			.474	1.226
	c_g			.821		.183	.060
	p_g				.998	.506	.200
	Mean	.923	.482	.192	.507	n = 55	
	S.D.	.314	1.253	.063	.201		

DISCRETE VERBAL ONLY CALIBRATION

ZGR1 (2/80)

		a_g	b_g	c_g	p_g	Mean	S.D.
ZGR1 (6/80)	a_g	.955				.905	.328
	b_g		.993			.469	1.225
	c_g			.842		.180	.049
	p_g				.998	.502	.202
	Mean	.912	.467	.182	.504	n = 55	
	S.D.	.333	1.243	.044	.204		

TABLE 23

Correlations and Summary Statistics for Item Parameters and Estimated Proportion Correct for the 25 Reading Comprehension Items of Section I of Form ZGR1

ALL VERBAL CALIBRATION

		ZGR1 (2/80)					
		a_g	b_g	c_g	p_g	Mean	S.D.
ZGR1 (6/80)	a_g	.918				.814	.175
	b_g		.992			-.039	.792
	c_g			.685		.167	.041
	p_g				.998	.585	.153
	Mean	.802	-.028	.171	.585		
	S.D.	.185	.831	.033	.156		

n = 25

READING COMPREHENSION ONLY CALIBRATION

		ZGR1 (2/80)					
		a_g	b_g	c_g	p_g	Mean	S.D.
ZGR1 (6/80)	a_g	.946				.932	.289
	b_g		.994			-.021	.773
	c_g			.709		.166	.039
	p_g				.998	.58[illegible]	.158
	Mean	.920	-.007	.166	.582		
	S.D.	.270	.823	.036	.164		

n = 25

TABLE 24

Correlations and Summary Statistics for Item Parameters and Estimated Proportion Correct for the 55 Quantitative Items of Section II of Form ZGR1

		ZGR1 (2/80)					
		a_g	b_g	c_g	p_g	Mean	S.D.
ZGR1 (6/80)	a_g	.969				.856	.398
	b_g		.996			.005	1.518
	c_g			.927		.183	.074
	p_g				.999	.576	.232
	Mean	.849	.020	.181	.573		
	S.D.	.391	1.517	.073	.231		

n = 55

to group differences and, showed little sample error, but were slightly sensitive to the reference set used for calibration, i.e., the difference between mean item difficulties, b_g, is greater in the verbal calibrations (.474 vs. .482) than the difference between mean item difficulties for the discrete verbal calibrations (.469 vs. .467). Note also, the corresponding differences in c_g calibrations (verbal calibrations, .183 vs. .192; discrete verbal calibrations, .180 vs. .182. The differences in p_g (.506 vs. .507 for verbal and .502 vs. .504 for discrete verbal) indicate that these differences compensate for each other. One can infer that these differences are probably due to error of estimation. Note that c_g exhibits the least temporal stability.

Table 23, which has the same format as Table 22, contains means, standard deviations, and correlations obtained for all 25 reading comprehension items in Section I of form ZGR1. Note that p_g is virtually insensitive to group differences or item calibration reference set. The consistency of the item response theory estimate of difficulty, b_g, however, is slightly imperfect. The most notable effect evident in Table 23 is sensitivity of a_g to homogeneity of item calibration set: The mean a_g for the 25 reading comprehension items is higher when these 25 items are calibrated with reading comprehension items only than when calibrated with all verbal items. Further discussion of homogeneity effects is deferred to the next section. The final point to note in Table 23 is the comparatively low correlations obtained between c_g estimates. This is due to the relative easiness of the reading comprehension items (b slightly below .0 as opposed to discrete verbal b of about .5). It is difficult to estimate c for easy items because of insufficient data at the lower asymptote.

Table 24 contains the means, standard deviations, and correlations obtained for the 55 quantitative items in Section II of form ZGR1. The high correlations for a_g and c_g and the overall stability of item parameter estimates are the notable features of this table.

Sensitivity of Item Parameter Estimates to Violations of Unidimensionality

Evidence indicating that verbal items are not homogeneous, i.e., that they measure more than one dimension, was presented in the sections of this report dealing with the factor analytic review, the violation of local independence, and item-ability regressions. In this section, the comparability of item parameter estimates based on calibration of heterogenous (all verbal) and homogeneous (discrete verbal only and reading comprehension only) item sets is assessed. Calibrations from all five administrations, ZGR1(6/80), ZGR1(2/80), K-ZGR2(12/79), K-ZGR3(4/80) and 3CGR1(6/80), are examined.

Table 25 contains the results for estimates of item discrimination (a_g). The results for discrete verbal items appear in the top half of the table, while the bottom half contains the results for reading comprehension items. The elements in the top rectangle of Table 25 are correlations between item discrimination estimates based on verbal and discrete verbal calibrations,

TABLE 25

Summary Statistics for and Correlations Between Parameter Estimates of Item Discrimination (a_g) Based on Sets of Homogeneous and Heterogenous Items

DISCRETE VERBAL ONLY

		ZGR1 (6/80)	ZGR1 (2/80)	K-ZGR2 (12/79)	K-ZGR3 (4/80)	CGR1 (6/80)	n	Mean	S.D.
ALL VERBAL	ZGR1 (6/80)	.969					108	.922	.316
	ZGR1 (2/80)		.975				102	.885	.344
	K-ZGR2 (12/79)			.984			102	.898	.343
	K-ZGR3 (4/80)				.975		102	.874	.336
	CGR1 (6/80)					.976	108	.954	.320
	n	108	102	102	102	108			
	Mean	.930	.881	.936	.876	.963			
	S.D.	.331	.357	.380	.344	.328			

READING COMPREHENSION ONLY

		ZGR1 (6/80)	ZGR1 (2/80)	K-ZGR2 (12/79)	K-ZGR3 (4/80)	CGR1 (6/80)	n	Mean	S.D.
ALL VERBAL	ZGR1 (6/80)	.800					47	.791	.200
	ZGR1 (2/80)		.889				45	.730	.237
	K-ZGR2 (12/79)			.926			45	.730	.245
	K-ZGR3 (4/80)				.904		45	.759	.285
	CGR1 (6/80)					.902	47	.761	.201
	n	47	45	45	45	47			
	Mean	.867	.837	.824	.848	.844			
	S.D.	.287	.338	.349	.324	.271			

while the correlations between estiamtes based on verbal and reading comprehension calibrations appear in the bottom rectangle.

Under the top rectangle are the number of items calibrated (n), means, and standard deviations of the a_g for the discrete verbal calibrations at each of the five administrations. To the right of the top rectangle are the summary statistics for the corresponding verbal calibrations of the discrete verbal items. Under the bottom rectangle are summary statistics for the five reading comprehension calibrations, while the corresponding summary statistics for the five verbal calibrations of the reading comprehension items appears to the right of this bottom rectangle.

Tables 26, 27, and 28 are identical in format to Table 25 and contain the results for item difficulty (b_g) estimates, estimates of the psuedo-guessing parameter or lower asymptote (c_g), and estimated proportion correct (p_g).

The correlations and means in Table 25 reveal that the discrete verbal and verbal calibrations produce considerably more similiar estimates of a_g than the reading comprehension and verbal calibrations. The discrete verbal - verbal correlations between a_g estimates range from .97 to .98, while the reading comprehension - verbal correlations range from .80 to .93. The mean differences between a_g estimates for the discrete verbal items ranges from .00 to .04, while the range of mean differences for reading comprehension items is .07 to .11. When the smaller standard deviations of a_g estimates for reading comprehension items are considered, the magnitude of the mean differences for these items appears even larger relative to the magnitude of the mean difference for discrete verbal items.

Also evident from Table 25, in each pair of calibrations, for both discrete verbal-verbal and reading comprehension-verbal, is the fact that the standard deviation for the a_g estimates based on the more homogeneous calibrations is higher. The mean standard deviations of a_g estimates for the discrete verbal items based on the discrete verbal calibrations and the verbal calibrations are .349 and .332, respectively. Similarly, the mean standard deviations of a_g estimates for the reading comprehension items based on the reading comprehension calibrations and the verbal calibrations are .315 and .237, respectively. As with the differences in mean estimates, the difference in mean standard deviations of a_g estimates is more extreme for reading comprehension items than for discrete verbal items.

Evidence pertaining to the comparability of item difficulty estimates (b_g) appears in Table 26. The correlations and means in this table reveal that the discrete verbal and verbal calibrations produce slightly more similiar estimates than the reading comprehension and verbal calibrations. For the discrete verbal items, the correlations all round to 1.00, while the mean differences range from .00 to .01. For the reading comprehension items, the correlations range from .98 to 1.00 and the mean differences in b_g range from .00 to .03. When compared to the results for the a_g estimates, the b_g estimates show much less sensitivity to homogeneity of item calibration set.

TABLE 26

Summary Statistics for and Correlations Between Parameter Estimates of Item Difficulty (b_g) Based on Sets of Homogeneous and Heterogenous Items

DISCRETE VERBAL ONLY

		ZGR1 (6/80)	ZGR1 (2/80)	K-ZGR2 (12/79)	K-ZGR3 (4/80)	CGR1 (6/80)	n	Mean	S.D.
	ZGR1 (6/80)	.998					108	.336	1.229
	ZGR1 (2/80)		.996				102	.330	1.222
ALL VERBAL	K-ZGR2 (12/79)			.998			102	.269	1.284
	K-ZGR3 (4/80)				.999		102	.259	1.302
	CGR1 (6/80)					.998	108	.361	1.143
	n	108	102	102	102	108			
	Mean	.334	.335	.255	.265	.366			
	S.D.	1.237	1.211	1.281	1.330	1.154			

READING COMPREHENSION ONLY

		ZGR1 (6/80)	ZGR1 (2/80)	K-ZGR2 (12/79)	K-ZGR3 (4/80)	CGR1 (6/80)	n	Mean	S.D.
	ZGR1 (6/80)	.993					47	.167	.952
	ZGR1 (2/80)		.994				45	.433	.978
ALL VERBAL	K-ZGR2 (12/79)			.996			45	.367	.959
	K-ZGR3 (4/80)				.978		45	.369	1.092
	CGR1 (6/80)					.995	47	.180	.954
	n	47	45	45	45	47			
	Mean	.162	.453	.387	.347	.152			
	S.D.	.950	.979	.981	1.060	.921			

TABLE 27

Summary Statistics for and Correlations Between Parameter Estimates of Lower Asymptote (c_g) Based on Sets of Homogeneous and Heterogenous Items

DISCRETE VERBAL ONLY

		ZGR1 (6/80)	ZGR1 (2/80)	K-ZGR2 (12/79)	K-ZGR3 (4/80)	CGR1 (6/80)	n	Mean	S.D.
ALL VERBAL	ZGR1 (6/80)	.897					108	.177	.054
	ZGR1 (2/80)		.767				102	.183	.053
	K-ZGR2 (12/79)			.874			102	.179	.049
	K-ZGR3 (4/80)				.940		102	.175	.040
	CGR1 (6/80)					.932	108	.181	.058
	n	108	102	102	102	108			
	Mean	.176	.180	.161	.173	.177			
	S.D.	.047	.040	.059	.040	.059			

READING COMPREHENSION ONLY

		ZGR1 (6/80)	ZGR1 (2/80)	K-ZGR2 (12/79)	K-ZGR3 (4/80)	CGR1 (6/80)	n	Mean	S.D.
ALL VERBAL	ZGR1 (6/80)	.658					47	.165	.043
	ZGR1 (2/80)		.844				45	.168	.031
	K-ZGR2 (12/79)			.800			45	.169	.033
	K-ZGR3 (4/80)				.550		45	.175	.065
	CGR1 (6/80)					.923	47	.164	.039
	n	47	45	45	45	47			
	Mean	.159	.168	.172	.168	.158			
	S.D.	.042	.034	.037	.037	.039			

TABLE 28

Summary Statistics for and Correlations Between Parameter Estimates of Proportion Correct (p_g) Based on Sets of Homogeneous and Heterogenous Items

DISCRETE VERBAL ONLY

		ZGR1 (6/80)	ZGR1 (2/80)	K-ZGR2 (12/79)	K-ZGR3 (4/80)	CGR1 (6/80)	n	Mean	S.D.
ALL VERBAL	ZGR1 (6/80)	.999					108	.523	.205
	ZGR1 (2/80)		.999				102	.529	.202
	K-ZGR2 (12/79)			.999			102	.539	.210
	K-ZGR3 (4/80)				.999		102	.537	.218
	CGR1 (6/80)					.999	108	.519	.202
	n	108	102	102	102	108			
	Mean	.520	.527	.531	.535	.515			
	S.D.	.208	.204	.210	.220	.204			

READING COMPREHENSION ONLY

		ZGR1 (6/80)	ZGR1 (2/80)	K-ZGR2 (12/79)	K-ZGR3 (4/80)	CGR1 (6/80)	n	Mean	S.D.
ALL VERBAL	ZGR1 (6/80)	.999					47	.553	.172
	ZGR1 (2/80)		.998				45	.513	.169
	K-ZGR2 (12/79)			.999			45	.522	.171
	K-ZGR3 (4/80)				.999		45	.520	.187
	CGR1 (6/80)					.998	47	.555	.167
	n	47	45	45	45	47			
	Mean	.551	.509	.522	.517	.554			
	S.D.	.177	.174	.178	.190	.174			

The results pertaining to the sensitivity of c_g estimates to homogeneity of item calibration set are portrayed in Table 27. With the exception of the discrete verbal items on form K-ZGR2, all mean differences in this table are all less than .01. Compared to those in Tables 25 and 26, correlations in Table 27 are low and more variable, reflecting difficulties inherent in obtaining stable estimates of c_g (Lord, 1975b).

Table 28 reveals that the similarity of p_g estimates based on heterogenous vs. homogeneous calibrations is very high. This high degree of similarity is evident for both discrete verbal items and reading comprehension items, as is reflected in the means and correlations in this table. An inference suggested by the results in Table 28 is that the observed data can be approximated equally as well by sets of heterogeneous items (all verbal) as by sets of homogeneous items (discrete verbal and reading comprehension). This inference was also suggested by the results discussed in the section on item-ability regressions.

Comparability of Ability Estimates Based on Homogenous and Heterogeneous Sets of Items

The review of factor analytic studies conducted on the GRE Aptitude Test led to a decision to separate verbal items into mutually exclusive sets of discrete verbal items and reading comprehension items because the evidence indicated that the items on the verbal scale were measuring two correlated factors. Consequently, all verbal items were calibrated at least twice, once with a set of homogeneous items of like type, e.g., discrete verbal or reading comprehension, and once with a set of heterogenous items comprised of both discrete verbal and reading comprehension items. This procedure produced three ability scores for each examinee: a verbal ability score based on all verbal items (θ_V), a discrete verbal ability score based on discrete verbal items (θ_{DV}), and a reading comprehension ability score based on reading comprehension items (θ_R).

If discrete verbal items and reading comprehension items were measuring the same attribute, then ability estimates based on each set of items should be very highly correlated. On the other hand, if these different sets of items were measuring distinct abilities, the expected correlation would not be as high. Table 29 provides evidence relevant to assessing whether the reading comprehension items and the discrete verbal items are measuring the same attribute. It contains correlations among θ_V, θ_{DV}, and θ_K for all four administrations.

It is clear in Table 29 that discrete verbal ability had a higher correlation with verbal ability than did reading comprehension ability, and that discrete verbal ability and reading comprehension ability were less correlated with each other than with verbal ability. The three correlations are .96 to .97 for discrete verbal ability and verbal ability, .86 to .89 for reading comprehension ability and verbal ability, and .73 to .77 for discrete verbal ability and reading comprehension ability. Since estimated θ has about the same reliability as the usual number-right test score, a correction for attenuation due to error of

TABLE 29

Correlations Among Ability Estimates for Verbal (V), Discrete Verbal (DV) and Reading Comprehension (R) Scales

Admin Date	Form															
	ZGR1				K-ZGR2				K-ZGR3				3CGR1			
12/79						DV	V	R								
					DV	1.000	.959	.726								
					V		1.000	.860								
					R			1.000								
					N = 3861											
2/80		DV	V	R												
	DV	1.000	.965	.764												
	V		1.000	.881												
	R			1.000												
	N = 3581															
4/80										DV	V	R				
									DV	1.000	.965	.766				
									V		1.000	.886				
									R			1.000				
									N = 4043							
6/80		DV	V	R										DV	V	R
	DV	1.000	.968	.746									DV	1.000	.970	.758
	V		1.000	.861									V		1.000	.863
	R			1.000									R			1.000
	N = 4351												N = 2579			

estimation is probably necessary. Assuming that this correction has little differential effect on the correlations, then the correlations in Table 29 indicate that discrete verbal ability and reading comprehension ability are distinct, highly correlated abilities.

Further evidence for the conclusion that reading comprehension ability and discrete verbal ability are distinct, highly correlated abilities is presented in Table 30, which contains correlations among proportion-correct true scores for verbal, discrete verbal, and reading comprehension abilities. Proportion-correct true score is obtained by substituting ability estimates into the test characteristic curve, which is a sum of the item characteristic curves for the items defining the test, and dividing the result, which is the number-correct true score, by the number of items in the test. Preférence for correlations of bounded difficulty parameters was one reason for examining proportion-correct true score.

The correlations in Table 30 present a range of .96 to .98 for the discrete verbal-verbal correlation, a range of .88 to .90 for the reading comprehension-verbal correlation, and a range of .73 to .80 for the discrete verbal-reading comprehension correlation. These latter results, like the results in Table 29, provide evidence for the existence of the two distinct, highly correlated reading comprehension and discrete verbal abilities.

The fourth column in Table 30 contains the correlations of the variable V* with the discrete verbal, verbal, and reading comprehension proportion-correct true scores. This variable, V*, is defined as the sum of the discrete verbal number-correct true score and the reading comprehension number-correct true score divided by the total number of items, i.e., V* is a weighted composite of the discrete verbal and reading comprehension proportion-correct true scores, where the weights are the number of discrete verbal items and the number of reading comprehension items, respectively.

The striking feature of the fourth columns in Table 30 is the close resemblance of the V* correlations to the verbal (V) correlations. For all five administrations, V and V* are virtually perfectly correlated, and their correlations with discrete verbal (DV) and reading comprehension (R) are almost identical. Hence, Table 30 provides evidence for thinking of the verbal true score dimension as a weighted composite of the discrete verbal and reading comprehension dimensions. Table 31 provides further support for this inference.

Table 31 contains means and standard deviations for the verbal (V), discrete verbal (DV), reading comprehension (R), and reconstructed verbal (V*) proportion-correct true scores for all five administrations. Note that the maximum difference between verbal (V) and reconstructed verbal (V*) means and standard deviations is .001, which provides further support for viewing verbal ability as a weighted composite of discrete verbal ability and reading comprehension ability.

TABLE 30

Correlations Among Proportion-Correct True Score Estimates for Verbal (V), Discrete Verbal (DV), Reading Comprehension (R) and Reconstructed Verbal (V*) Scales

Admin Date	Form		DV	V	R	V*
12/79	K-ZGR2	DV	1.000	.963	.734	.968
		V		1.000	.879	.996
		R			1.000	.882
		N = 3861				
2/80	ZGR1	DV	1.000	.961	.758	.968
		V		1.000	.899	.996
		R			1.000	.898
		N = 3581				
4/80	K-ZGR3	DV	1.000	.962	.768	.971
		V		1.000	.902	.995
		R			1.000	.899
		N = 4043				
6/80	ZGR1	DV	1.000	.971	.775	.971
		V		1.000	.901	.999
		R			1.000	.903
		N = 4351				
6/80	3CGR1	DV	1.000	.980	.798	.982
		V		1.000	.898	.999
		R			1.000	.898
		N = 2579				

TABLE 31

Summary Statistics for Verbal (V), Discrete Verbal (DV), Reading Comprehension (R), and Reconstructed Verbal (V*) Proportion-Correct True Score Estimates

Form		DV	R	V	V*
ZGR1 (6/80)	Mean	.518	.615	.549	.548
	S.D.	.152	.194	.155	.156
ZGR1 (2/80)	Mean	.523	.624	.554	.555
	S.D.	.151	.195	.154	.155
K-ZGR2 (12/79)	Mean	.560	.656	.590	.590
	S.D.	.153	.185	.152	.152
K-ZGR3 (4/80)	Mean	.532	.631	.562	.563
	S.D.	.142	.175	.144	.144
3CGR1 (6/80)	Mean	.547	.570	.555	.554
	S.D.	.165	.163	.157	.157

Further evidence pertaining to the dimensionality of the verbal items is also presented in Table 32, which contains correlations among observed scores, with and without correction for attenuation due to measurement error, on the verbal item types for four distinct samples of examinees who took one of these four forms in June, 1980: $\mathrm{ZGR1}_{C47}$, $\mathrm{ZGR1}_{C48}$, $\mathrm{3CGR1}_{C41}$ and $\mathrm{3CGR1}_{C42}$. The elements on the main diagonals of the four correlation matrices in Table 32 are reliability estimates. An adaptation of Kuder-Richardson formula 20 (KR-20) for formula-scored tests (Dressel, 1940) produced the reliability estimates for sentence completions, analogies, antonyms, and reading comprehension. These four modified KR-20 estimates were used to estimate the reliability for the verbal scale via the formula

$$(4) \qquad \mathrm{Rel}_v = 1 - \sum_{i=1}^{4} \mathrm{Var}_i(1-\mathrm{Rel}_i)/\mathrm{Var}_v ,$$

where Rel_v and Var_v are the reliability and variance, respectively, of the verbal scale, and Var_i and Rel_i are the variance and modified KR-20 reliability estimate of the ith scale, where i is either one of the three discrete verbal item types or the reading comprehension item type. To obtain the reliability estimate for the discrete verbal scale, the above formula is used with the three discrete verbal item type variances and reliabilities and the discrete verbal variance.

The elements to the left of the main diagonal are observed score correlations, while the entries to the right are the same correlations corrected for attenuation. Note that part-total correlations, such as the five correlations with verbal score, were not corrected for attenuation.

The disattenuated correlations between discrete verbal and reading comprehension are of primary interest. Since the reliabilities used to correct the observed score correlations for attenuation are estimates of item homogeneity, the reliabilities reported on the diagonals in Table 32 are probably underestimates. Hence, the disattenuated correlations in this table can be viewed as overestimates of the true score correlations among the verbal item types. The correlations between estimated proportion correct true scores for discrete verbal and reading comprehension on the June 1980 administrations of forms ZGR1 ($r = .775$) and 3CGR1 ($r = .798$), reported in Table 30, fall between the upper bound disattenuated correlations and the observed score correlations reported in Table 32, providing further evidence for the hypothesis that the verbal ability measured by the GRE Aptitude Test is composed of two distinct, highly correlated reading comprehension and discrete verbal abilities.

Table 32

Correlations Among Verbal Item Types With and Without Correction For Attenuation*

ZGR1

		C_{47}(N = 2,480)							C_{48}(N = 2,485)						Number of Items
		1	2	3	4	5	6		1	2	3	4	5	6	
1	Verbal	.929						1	.934						80
2	Discrete Verbal	.957	.896				.811	2	.956	.901				.806	55
3	Sentence Compl.	.877	.888	.759	.930	.880	.864	3	.882	.898	.765	.946	.903	.852	17
4	Analogies	.854	.895	.693	.732	.978	.795	4	.859	.895	.710	.736	.969	.807	18
5	Antonyms	.831	.894	.669	.730	.761	.710	5	.863	.901	.697	.734	.779	.705	20
6	Reading Comp.	.882	.710	.696	.629	.573	.855	6	.886	.713	.695	.645	.580	.869	25

3CGR1

		C_{41} (N = 1,485)							C_{42} (N = 1,495)						Number of Items
		1	2	3	4	5	6		1	2	3	4	5	6	
1	Verbal	.929						1	.931						75
2	Discrete Verbal	.974	.911				.858	2	.975	.913				.872	53
3	Sentence Compl.	.845	.845	.718	.894	.863	.899	3	.839	.841	.709	.917	.857	.909	13
4	Analogies	.863	.886	.653	.743	.909	.847	4	.873	.895	.664	.740	.939	.869	18
5	Antonyms	.889	.927	.677	.726	.858	.768	5	.897	.931	.670	.750	.863	.795	22
6	Reading Comp.	.864	.728	.677	.649	.632	.790	6	.874	.745	.684	.668	.660	.799	22

*Upper triangle has correlations corrected for attenuation;
diagonal has reliability estimates;
lower triangle has uncorrected correlations.

IRT EQUATING:

COMPARABILITY WITH LINEAR AND EQUIPERCENTILE EQUATING

In preceding sections of this report, the reasonableness of the assumptions of item response theory for the GRE Aptitude Test has been assessed. Evidence has been presented that, to some extent, the assumption of unidimensionality is violated within each section of the Aptitude Test. Despite these violations, the analysis of item-ability regressions indicated that, for the verbal items and two of the three analytical item types (logical diagrams and analytical reasoning), the three-parameter logistic model fit the data well. The quantitative item types, particularly quantitative comparison items, and the analysis of explanations items were fit less well by the model. Some quantitative comparison and analysis of explanations items showed local instances of an inverse relationship between the probability of responding correctly to the item and estimated theta (i.e., nonmonotonicity). Nonetheless, IRT-based equating might well be robust to violations of these assumptions. This section will compare a variety of equatings for three forms of the GRE Aptitude Test. The equating methods will be described, the equating plan will be outlined, and the results of the various equatings will be presented, compared, and analyzed.

Equating Methods

In practice, despite efforts by test development experts, two forms of the GRE Aptitude Test cannot be expected to be of precisely equal difficulty. Since it is inherently unfair to compare without adjustment the raw scores of examinees who take two tests that differ in difficulty, equating procedures have been developed to transform scores from different test forms to a single scale. These equating procedures each consist of two parts, a data collection design and an analytical method to determine the appropriate transformation.

There are three basic designs for data collection: single group, equivalent group, and anchor test (Lord, 1975a). Equatings considered in this study are based on the latter two designs. In the equivalent-group design, the old form (form already on scale) and new form (form to be scaled) are administered to random or otherwise equivalent samples from the same populations. In practice this is done through a procedure known as spiralling (Conrad, Trismen, & Miller, 1977). Test books are packaged alternating the old and new forms and then administered within each test center so that half the examinees within each test center take each form. The anchor-test design is one in which one form of the test is administered to one group, another form to another group, and a common anchor test to both groups. The anchor test allows the equating transformation to take the difference in abilities of the two groups into account; the equivalent-group method depends on spiralling to minimize ability differences.

Three major analytical methods to determine equating transformations were used in this research: equipercentile, linear, and item response theory based true score equating. In equipercentile equating, a transformation is chosen such that scores from the two tests will be considered equated if they correspond to the same percentile rank in some group of examinees. For linear equating, the chosen transformation is such that scores from the two tests will be considered equated if they correspond to the same number of standard deviations from the mean in some group of examinees. The transformation chosen for item response theory based equating is such that true scores from the two tests will be considered equated if they correspond to the same estimated theta (see Lord, 1980a, chapter 13.5 for a more complete description of item response theory based true score equating).

Nine variants of item response theory based equating were performed in this research. These variants differ along three dimensions: (a) the data collection design: equivalent group or anchor test; (b) the item parameter linking procedure; and (c) the composition of the item sets used in the LOGIST calibrations. For the equivalent-group design, the separate calibrations for the old and new forms are assumed to be on the same scale based on group equivalence, or the items in the new form appeared in an experimental section of the old form and were calibrated in a single LOGIST run with the old form. For the anchor-test design, the parameter estimates were either linked by the Lord-Stocking robust procedure (further divided into number of links to the base scale: either one or two) or were not linked. Three variants of the composition of the item sets used in the LOGIST calibrations were investigated: both old and new forms had a single calibration per form of heterogeneous item types; the old form had a heterogeneous calibration, but the new form had two separate homogeneous calibrations; and both the old and new forms had two homogeneous calibrations per form. Not all possible combinations of these dimensions were used in this research.

Table 33 presents a concise description of the nine IRT equating variants studied in this research and indicates designations (to be used through the rest of this report) for each variant. Tables 33, 34, and 35 indicate which equating variants were used (respectively) for the verbal, quantitative, and analytical sections. Table 37 describes the three non-IRT equating variants. Tables 38, 39, and 40 indicate which of these variants were used for the verbal, quantitative, and analytical sections of each form.

The equating variant designations given in Tables 32 and 36 follow a straightforward pattern. The first character is the designation (I, E, or L) indicates the equating method is IRT, Equipercentile, or Linear. The second character (E or A) indicates the general data collection design, Equivalent group or Anchor test. The IRT equating variants are designated with three or four characters. The third character (S, P, L, or W) provides information about the linking of item parameter scales: separate calibrations whose scale equivalence is assumed based on Spiralling, item parameters Precalibrated in the variable section of the old form, item parameter scales linked using the Lord-Stocking robust linking procedure, or equating Without linking item parameters (Lord, 1981). The fourth character (V, H, or 2) either indicates the composition of item sets used in parameter estimation: a heterogeneous, all Verbal items, single calibration for the old form and two homogeneous, reading comprenension and discrete verbal separately, calibrations for the new form; two Homogeneous calibrations for both the old and the new form; or, in the case of the IAL2 equating, that there were 2 links in the chain to put the item parameter estimates on scale.

Table 33

Variants of IRT Equating and Their Designations

	Data Collection Design				
	Equivalent Group		Anchor Test		
Composition of item sets used in parameter estimation*	Separate calibrations of operational items in old and new forms assumed to be on scale based on group equivalence	All operational items in new form precalibrated in variable section of old form	Lord-Stocking robust linking procedure Number of links to base scale		Equating without linking item parameters (Lord, 1981)
			1	2	
Heterogeneous for old and new forms	IES	IEP	IAL	IAL2	IAW
Heterogeneous for old form; homogeneous for new form	IESV	**	IALV	**	**
Homogeneous for old and new forms	IESH	**	IALH	**	**

*The composition of item sets used in parameter estimation was varied only for verbal, for which discrete verbal items and reading comprehension items were calibrated separately in some analyses.

**These variants were not studied in this research.

Table 34

Verbal Equatings

IRT Variants and Forms Analyzed

	Data Collection Design				
	Equivalent Group		Anchor Test		
Composition of item sets used in parameter estimation*	Separate calibrations of operational items in old and new forms assumed to be on scale based on group equivalence	All operational items in new form precalibrated in variable section of old form	Lord-Stocking robust linking procedure Number of links to base scale		Equating without linking item parameters (Lord, 1981)
			1	2	
Heterogeneous for old and new forms	3CGR1	3CGR1	ZGR1 K-ZGR2 K-ZGR3	K-ZGR3	K-ZGR2 K-ZGR3
Heterogeneous for old form; homogeneous for new form	3CGR1	**	ZGR1 K-ZGR2 K-ZGR3	**	**
Homogeneous for old and new forms	3CGR1	**	ZGR1 K-ZGR2 K-ZGR3	**	**

**These variants were not studied in this research.

Table 35

Quantitative Equatings

IRT Variants and Forms Analyzed

	Data Collection Design				
	Equivalent Group		Anchor Test		
Composition of item sets used in parameter estimation	Separate calibrations of operational items in old and new forms assumed to be on scale based on group equivalence	All operational items in new form precalibrated in variable section of old form	Lord-Stocking robust linking procedure Number of links to base scale 1	 2	Equating without linking item parameters (Lord, 1981)
Heterogeneous for old and new forms	3CGR1	3CGR1	ZGR1 K-ZGR2 K-ZGR3	K-ZGR3	**
Heterogeneous for old form; homogeneous for new form	**	**	**	**	**
Homogeneous for old and new forms	**	**	**	**	**

**These variants were not studied in this research.

Table 36

Analytical Equatings

	Data Collection Design				
	Equivalent Group		Anchor Test		
Composition of item sets used in parameter estimation	Separate calibrations of operational items in old and new forms assumed to be on scale based on group equivalence	All operational items in new form precalibrated in variable section of old form	Lord-Stocking robust linking procedure Number of links to base scale		Equating without linking item parameters (Lord, 1981)
			1	2	
Heterogeneous for old and new forms	3CGR1	3CGR1	**	**	**
Heterogeneous for old form; homogeneous for new form	**	**	**	**	**
Homogeneous for old and new forms	**	**	**	**	**

**These variants were not studied in this research.

Table 37

Variants of Non-IRT Equating

and Their Designations

	Data Collection Design	
Method	Equivalent Group	Anchor Test
Equipercentile	EE	**
Linear	LE	LA

Table 38

Verbal Equatings

Non-IRT Variants and Forms Analyzed

	Data Collection Design	
Method	Equivalent Group	Anchor Test
Equipercentile	3CGR1 K-ZGR3	**
Linear	3CGR1 K-ZGR3	K-ZGR2

Table 39

Quantitative Equatings

Non-IRT Variants and Forms Analyzed

	Data Collection Design	
Method	Equivalent Group	Anchor Test
Equipercentile	3CGR1	**
Linear	3CGR1 K-ZGR3*	K-ZGR2

*Equated through a combination of single-group and equivalent-group designs; see text in equating plan section.

**This variant was not studied in this research.

Table 40

Analytical Equatings

Non-IRT Variants and Forms Analyzed

	Data Collection Design	
Method	Equivalent Group	Anchor Test
Equipercentile	3CGR1	**
Linear	3CGR1	K-ZGR2

**This variant was not studied in this research.

Equating Plan

All IRT equatings used form ZGR1 as the old form. Parameter estimates for the old form ZGR1 items were based on the June 1980 administration, with the exception of the IAW method which used data from the February 1980 administration. The linear and equipercentile equating for form 3CGR1 also used form ZGR1 administered in June 1980 as the old form. The verbal linear and equipercentile equatings of form K-ZGR3 used Form ZGR1 administered in December 1977 as the old form. The quantitative linear equating of K-ZGR3 was complicated by the changing of one item. The quantitative section of form K-ZGR3 was originally equated to form ZGR1 administered in December 1977 using the equivalent-group design. When one item was changed, the unchanged items were used in equating to the original, prechange form using data from April 1979, and then the total quantitative section including the revised item was equated to the 54 unchanged items using data from the April 1980 administration.

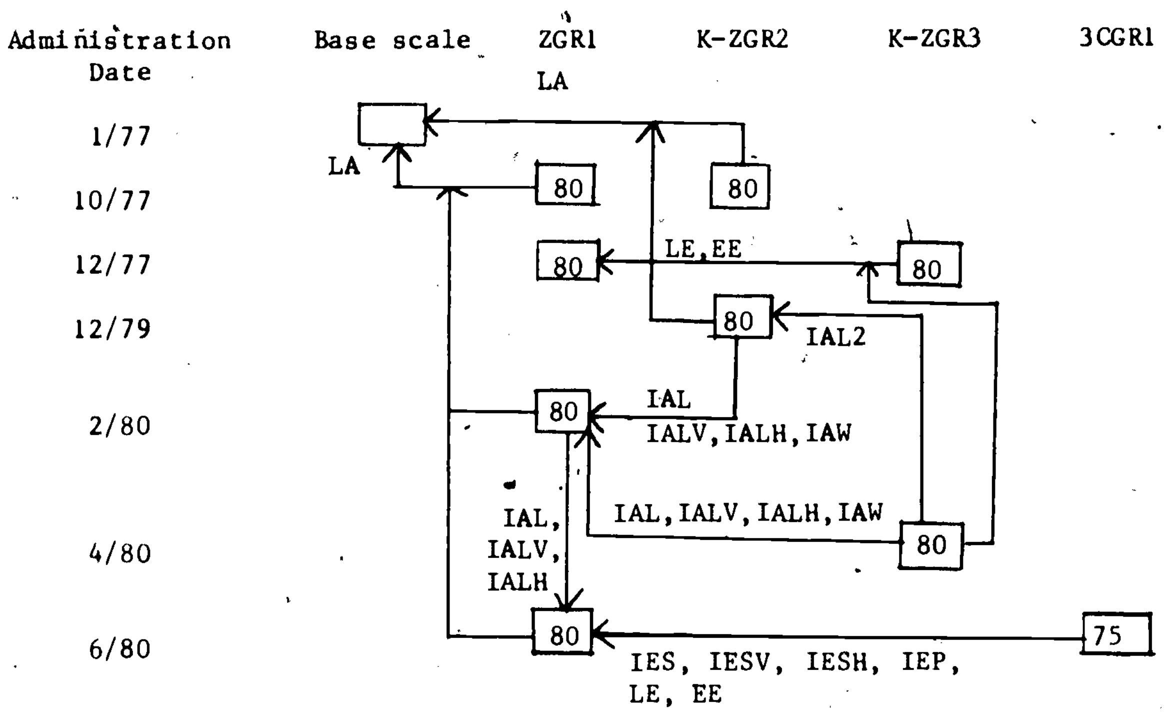
Figures 13, 14, and 15 present the equating plans for the verbal, quantitative, and analytical sections. Although, in the most obvious sense, ZGR1 administered in June 1980 (or February 1980 for the IAW equatings) is the old form for the IRT equatings (that is, the item parameters estimated from that administration's data were used), it is the item parameter scale linking (with the exception of the IAW method) that is most analogous to the equating links in linear or equipercentile equating plans. It is during these links that statistical error and bias can enter the equating system. The numbers in the boxes in Figures 13, 14, and 15 indicate the numbers of items in the operational section.

Judging the Adequacy of Equatings

Unfortunately, there is no unarguable objective criterion available to judge the adequacy of the equatings in this research. It is inappropriate to use the linear or equipercentile equatings as a criterion or method, particularly since (with the possible exception of the quantitative section in form 3CGR1) the assumptions upon which the linear and equipercentile methods are based are violated. As we have little evidence concerning the robustness of IRT equating to violations of its assumptions, we also have little evidence concerning the robustness of most of the classical methods (see, however, Marco, Petersen, & Stewart, 1979, and Petersen, Marco, & Stewart, in press, for a detailed analysis of the robustness of many anchor-test design methods). Further consideration of the assumptions of the equating variants used in this study, evidence concerning the violation of these assumptions, and interpretation of the equating results based on this evidence will be presented in the discussion section of this chapter.

Figure 13

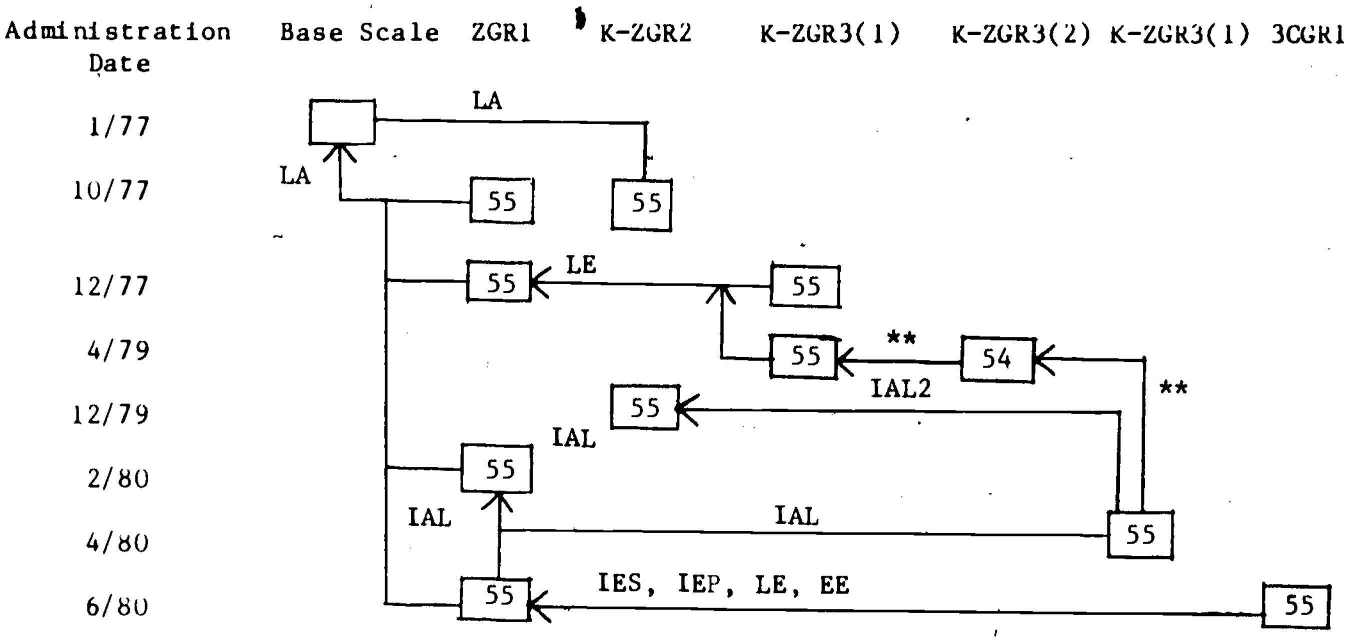
Equating Plan for Verbal Scales*



*The four administrations of form ZGR1, two administrations of form K-ZGR2, and two administrations of form K-ZGR3 are each assumed to be intraequated by virtue of the respective identity of their items.

Figure 14

Equating Plan for Quantitative Scales*

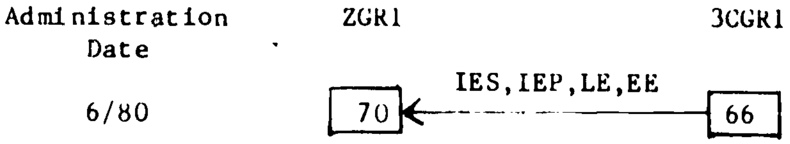


*The four administrations of form ZGR1, two administrations of form K-ZGR2, and two administrations of form K-ZGR3(1) are each assumed to be intraequated by virtue of the respective identity of their items.

**see text

Figure 15

Equating Plan for Analytical Scales



Results

Verbal equatings. Table 41 presents means, standard deviations, and skewnesses based on the various verbal equatings. Two factors went into the computation of these summary statistics: the relationship between raw and scaled score, as produced by the various equatings, and a frequency distribution of raw scores. This frequency distribution is simply a convenient vehicle for converting the vectors of scaled scores into the more easily interpretable, scalar, summary statistics presented. Any reasonable distribution would have been appropriate. The distributions used were based on the groups of examinees who took each of the forms when they were first administered. The equating tables and frequency distributions used to compute Table 41 are presented in Appendix A.

It should be noted that the means and standard deviations for the linear and equipercentile equatings based on the equivalent-group design are virtually identical. This is to be expected as they are based on identical data and the linear equating sets the first two moments of the old and new form distributions equal and the equipercentile equating sets all moments of the two distributions equal. Since only five significant digits were retained in the computations, minor differences due to small losses in accuracy in the computation of the standard deviations are noticeable.

Figures 16, 17, 18, and 19 plot the various equatings for the verbal sections of, respectively, forms ZGR1, K-ZGR2, K-ZGR3, and 3CGR1. This type of plot tends to point out the similarities between equatings more than the differences. A residuals plot is often more informative. In such a plot the difference between each equating and a comparison equating is plotted against raw score. Figures 20, 21, 22, and 23 are residuals plots using the IEP or IAL equating as the comparison, whichever is available.

Quantitative equatings. Table 42 was computed in the same way that Table 41 was computed and compares the various quantitative equatings. The equating tables and frequency distributions used to compute Table 42 are presented in Appendix A. Figures 24, 25, 26, and 27 are plots of the various quantitative equatings for forms ZGR1, K-ZGR2, K-ZGR3, and 3CGR1, respectively. Figures 28, 29, 30, and 31 are residuals plots using the IEP or IAL (whichever is available) equating as the comparison.

Analytical equatings. Table 43 presents the means, standard deviations, and skewnesses based on the analytical equatings of form 3CGR1. The equating tables and frequency distributions used to compute Table 43 are presented in Appendix A. Figures 32 and 33 are, respectively, a plot of the equatings and a residuals plot (using the IEP equating as the comparison) for the analytical section of form 3CGR1.

Discussion of Equatings

Lord (1980a, chapter 13) states that two tests cannot be equated unless they are perfectly reliable or strictly parallel. The first case

Table 41

Verbal Equatings
Means, Standard Deviations, and Skewnesses[a]

Equating Variant	Forms											
	3CGR1			K-ZGR3			K-ZGR2			ZGR1		
	Mean	S.D.	Skew	Mean	S.D.	Skew	Mean	S.D.	Skew	Mean	S.D.	Skew
IES	473.27	125.14	.14	*	*	*	*	*	*	*	*	*
IESV	475.80	123.39	.13	*	*	*	*	*	*	*	*	*
IESH	473.39	126.51	.15	*	*	*	*	*	*	*	*	*
IEP	473.81	125.47	.18	*	*	*	*	*	*	*	*	*
IAL	*	*	*	504.93	122.19	.08	496.68	125.14	.05	500.81	128.06	-.02
IALV	*	*	*	506.26	119.40	.12	500.46	120.12	.04	502.98	124.65	.02
IALH	*	*	*	504.54	122.58	.11	498.66	123.30	.05	501.26	127.78	.02
IAL2	*	*	*	504.22	122.13	.08	*	*	*	*	*	*
IAW	*	*	*	504.66	123.23	.14	503.18	125.66	.08	*	*	*
EE	473.29	123.30	.20	507.70	124.23	.03	*	*	*	*	*	*
LE	473.29	123.35	.10	507.70	124.20	.02	*	*	*	*	*	*
LA	*	*	*	*	*	*	502.16	126.26	-.01	501.69	126.75	.02

[a] The cells in this table in which asterisks appear represent equatings that were not carried out in this study.

Figure 16

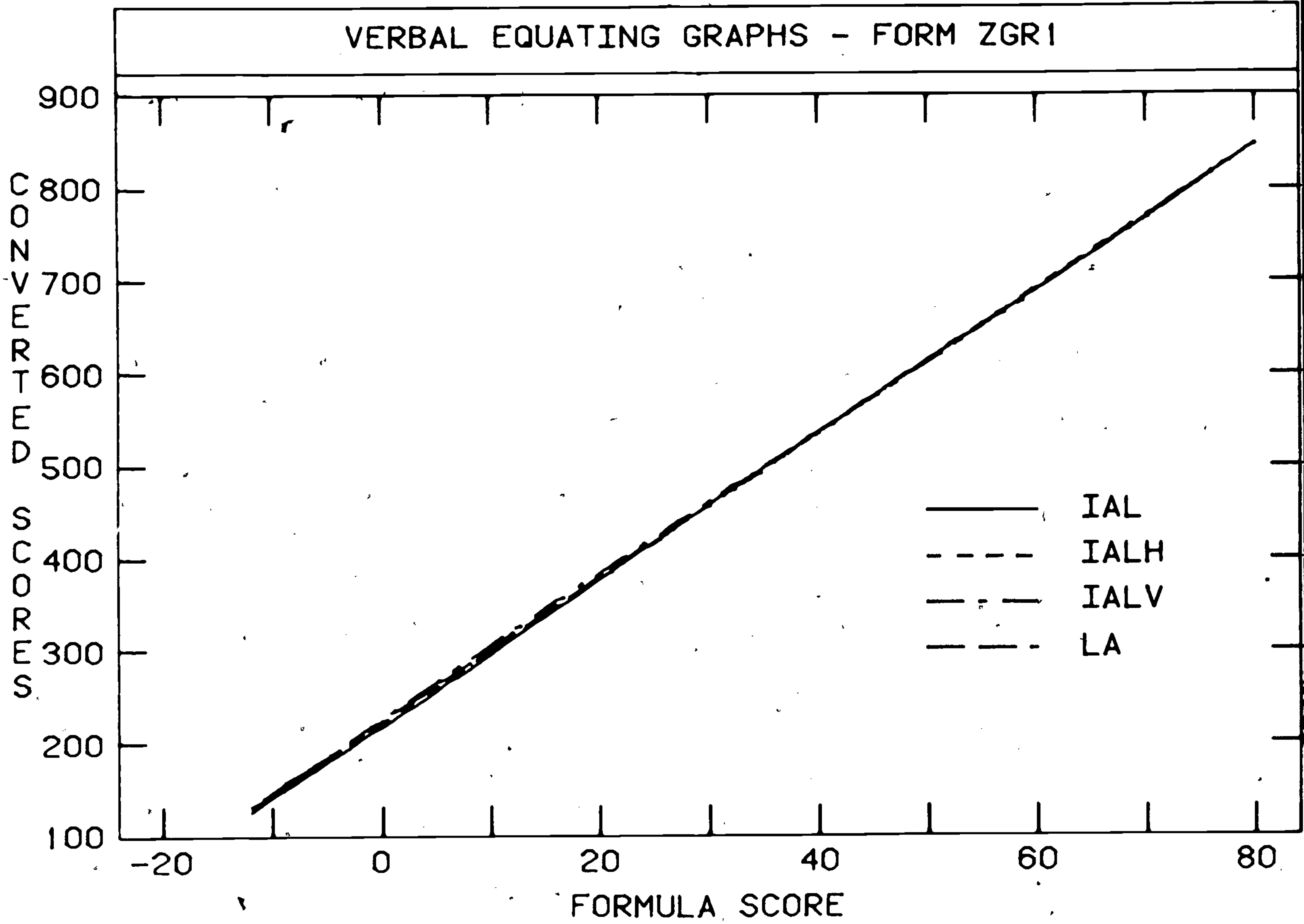

VERBAL EQUATING GRAPHS - FORM ZGR1
CONVERTED SCORES
900
800
700
600
500
400
300
200
100
-20
0
20
40
60
80
FORMULA SCORE
IAL
IALH
IALV
LA

Figure 17

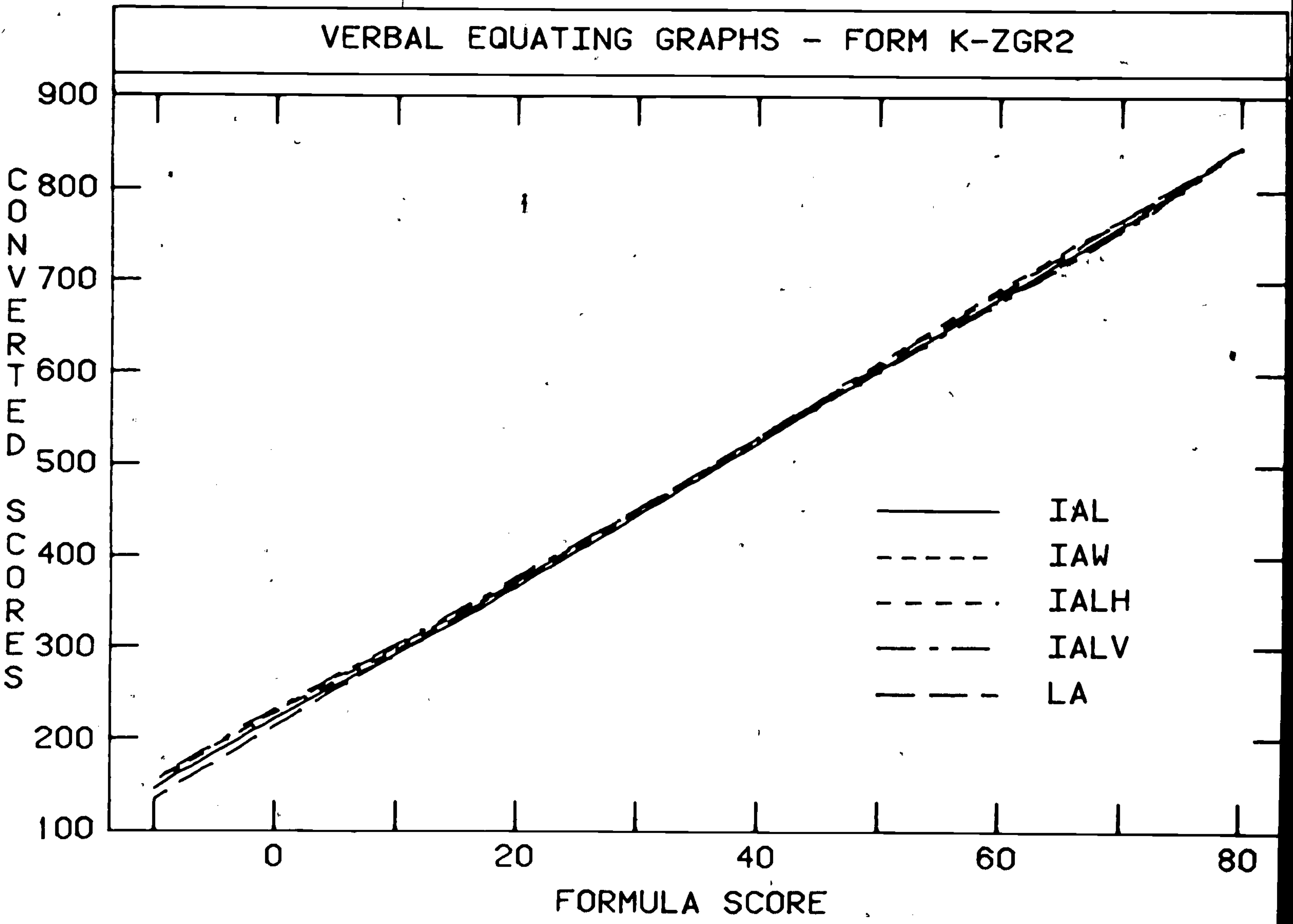
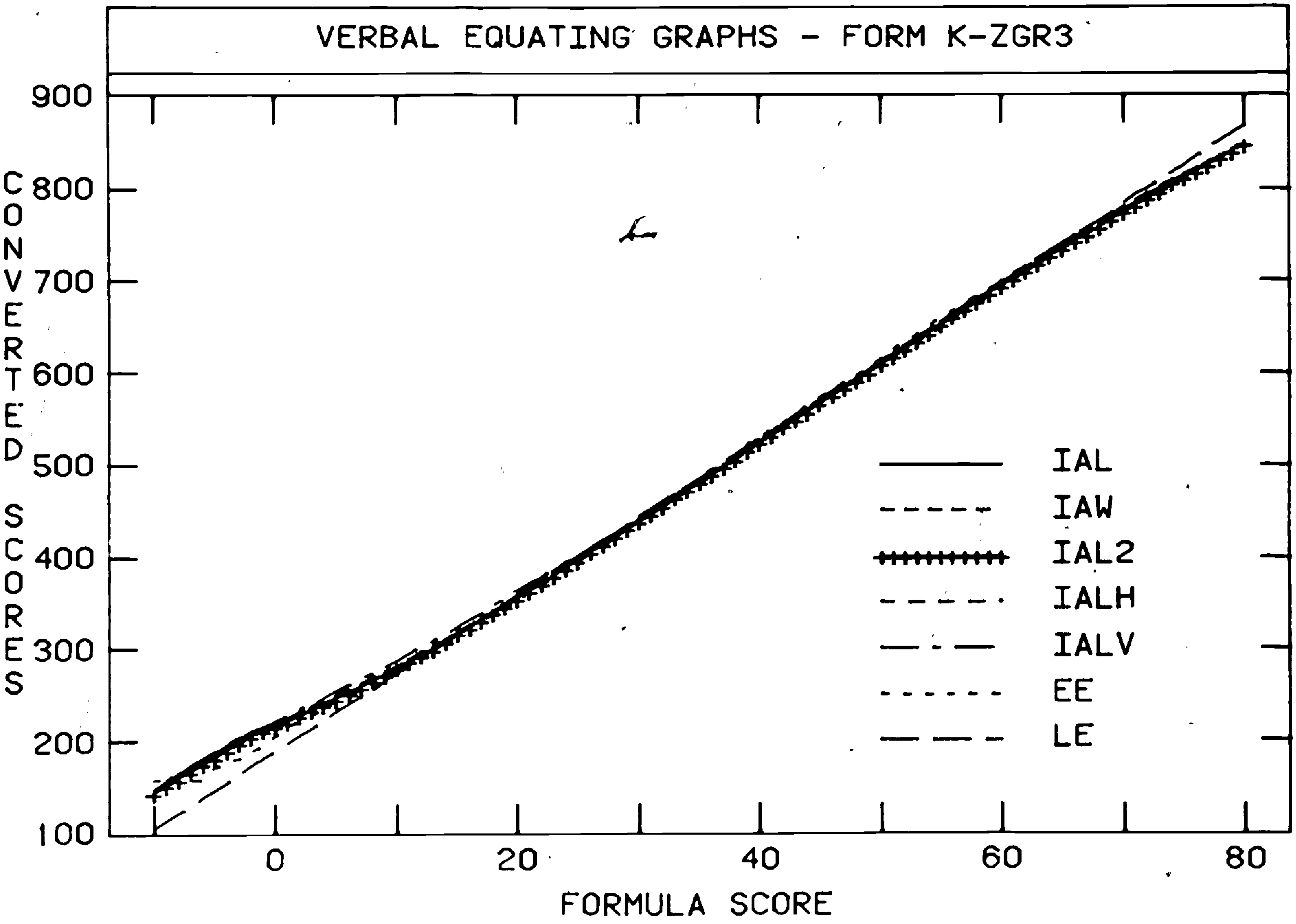

VERBAL EQUATING GRAPHS - FORM K-ZGR2
CONVERTED SCORES
900
800
700
600
500
400
300
200
100
0
20
40
60
80
FORMULA SCORE
IAL
IAW
IALH
IALV
LA

Figure 18



VERBAL EQUATING GRAPHS - FORM K-ZGR3
CONVERTED SCORES
900
800
700
600
500
400
300
200
100
0
20
40
60
80
FORMULA SCORE
IAL
IAW
IAL2
IALH
IALV
EE
LE

Figure 19

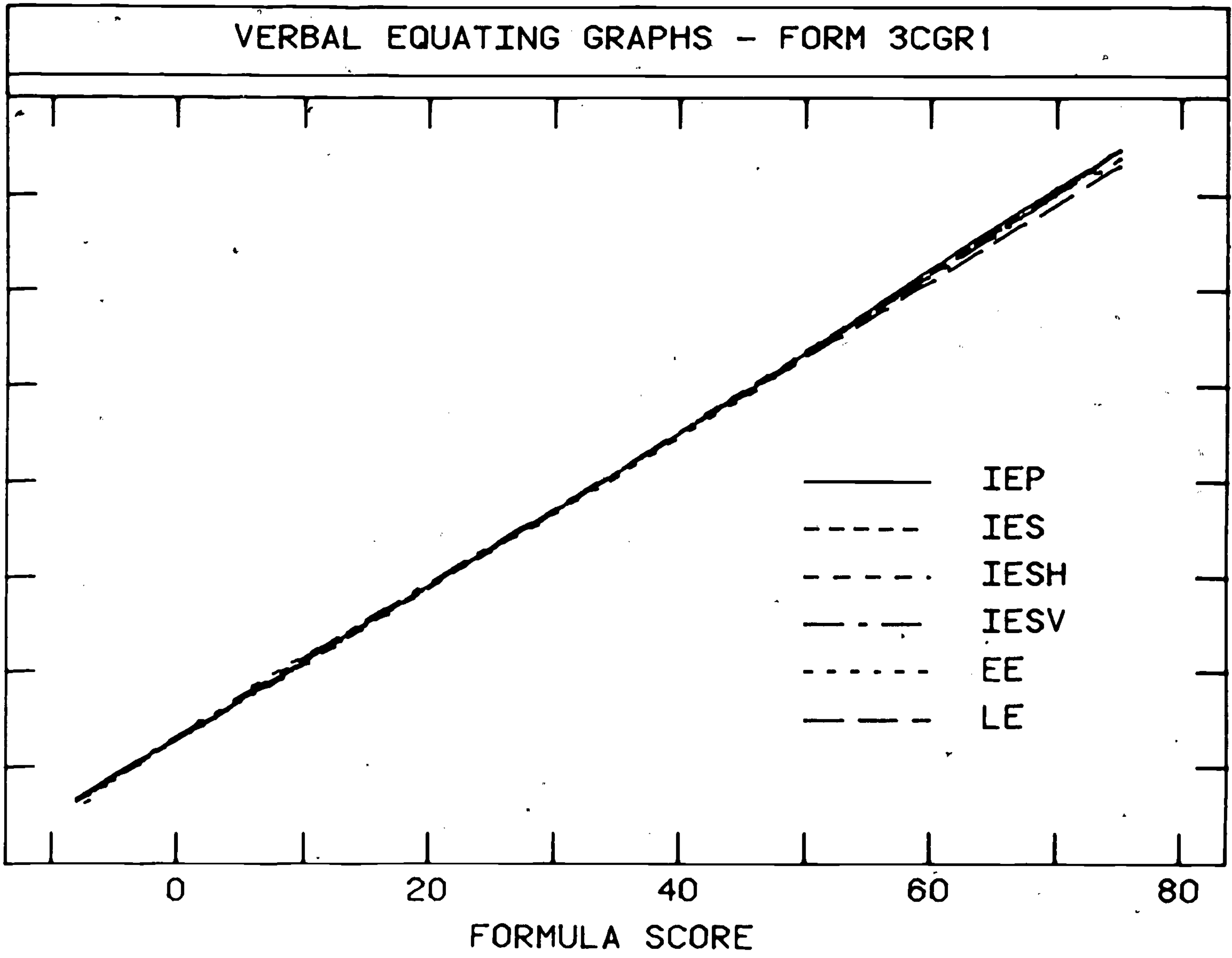
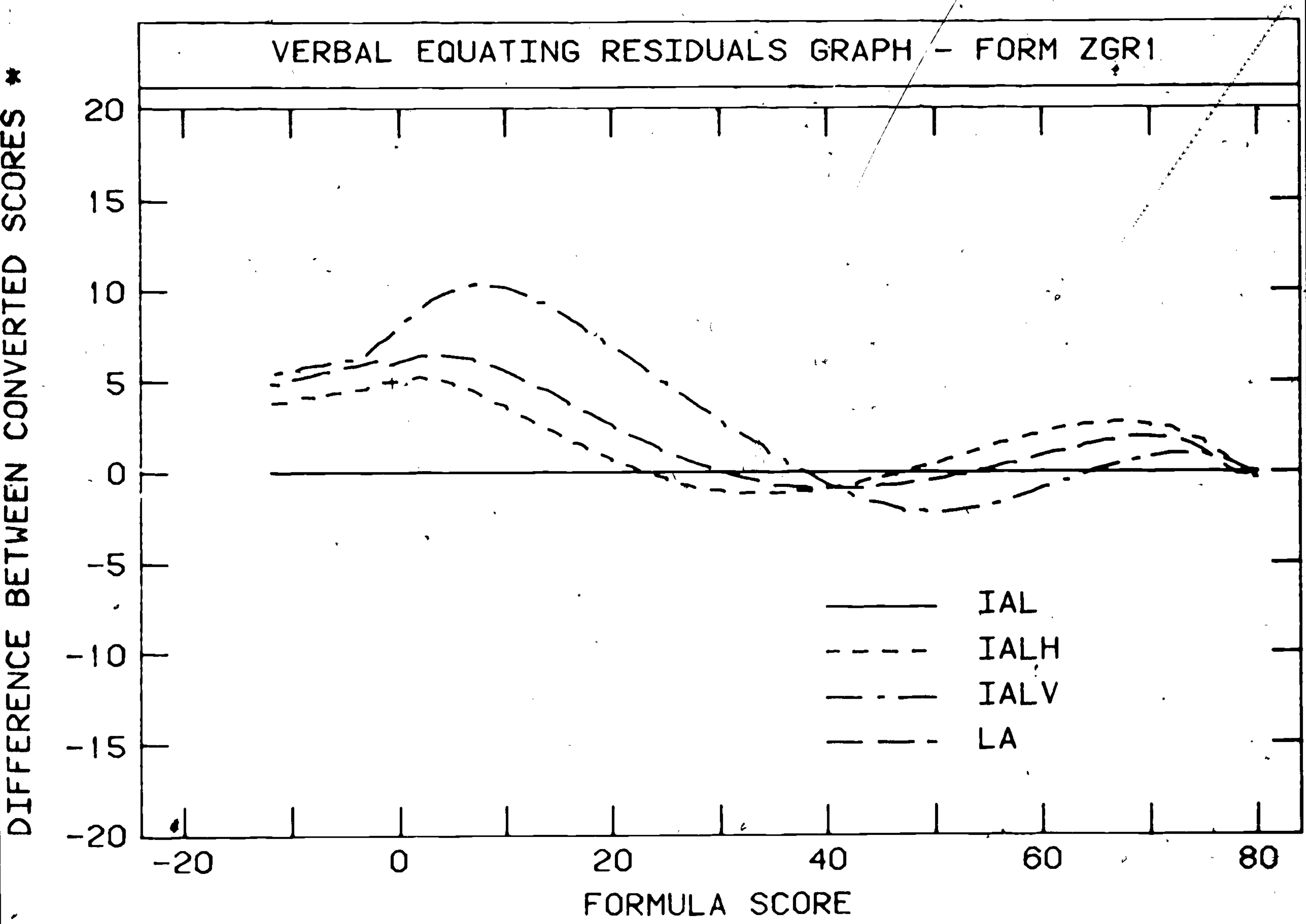
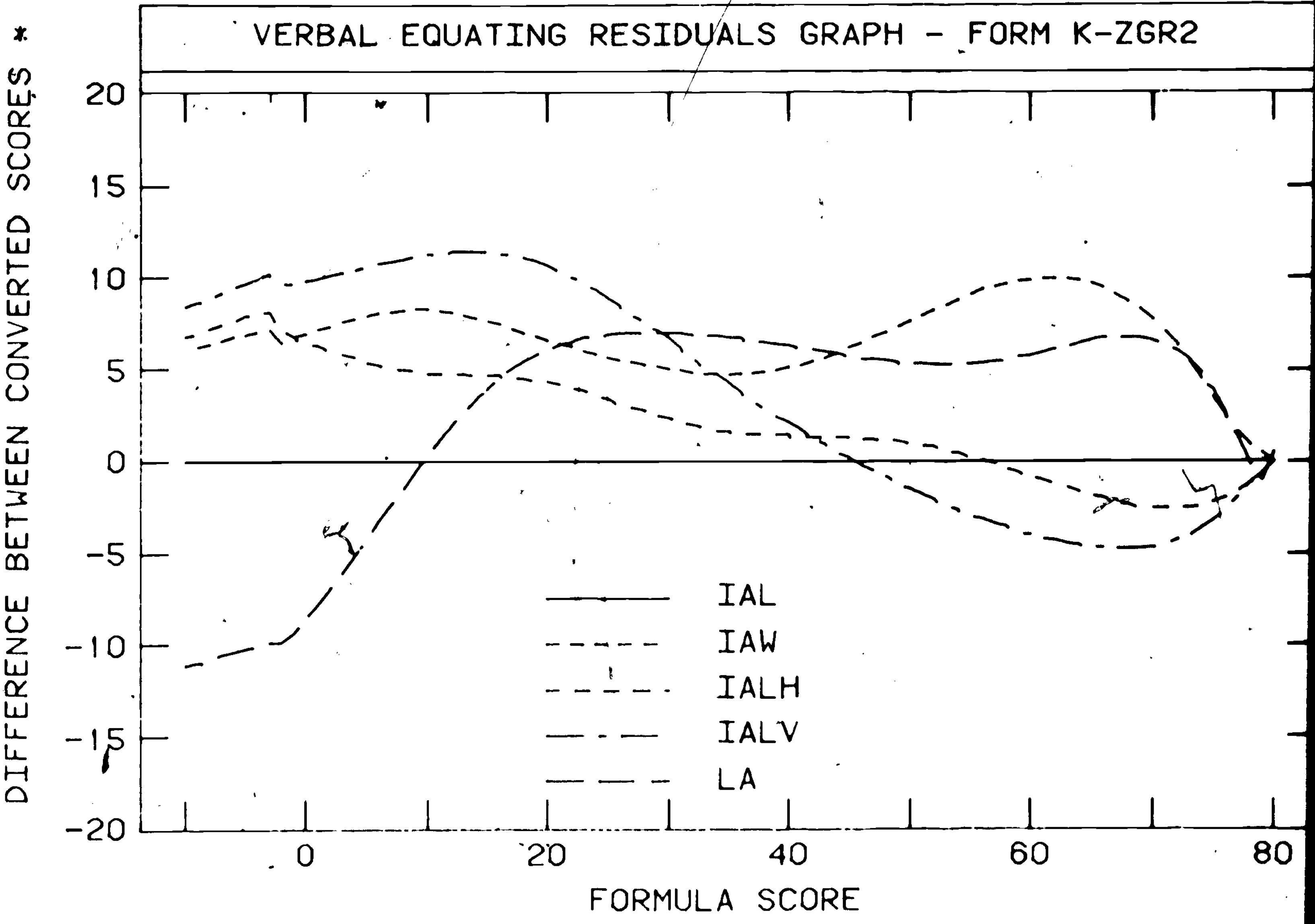

VERBAL EQUATING GRAPHS - FORM 3CGR1
CONVERTED SCORES
900
800
700
600
500
400
300
200
100
0
20
40
60
80
FORMULA SCORE
IEP
IES
IESH
IESV
EE
LE

Figure 20



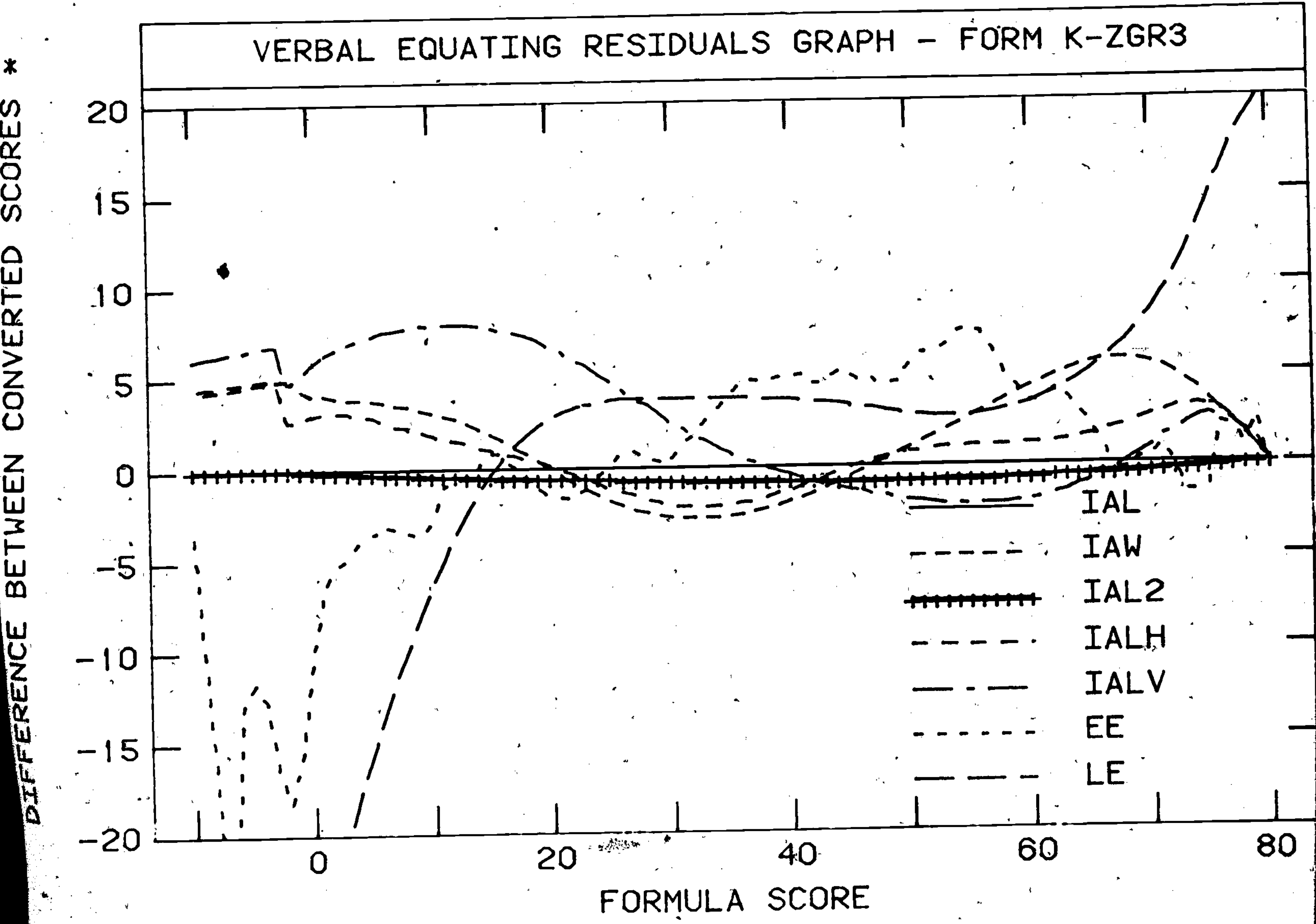
* SEE TEXT

Figure 21



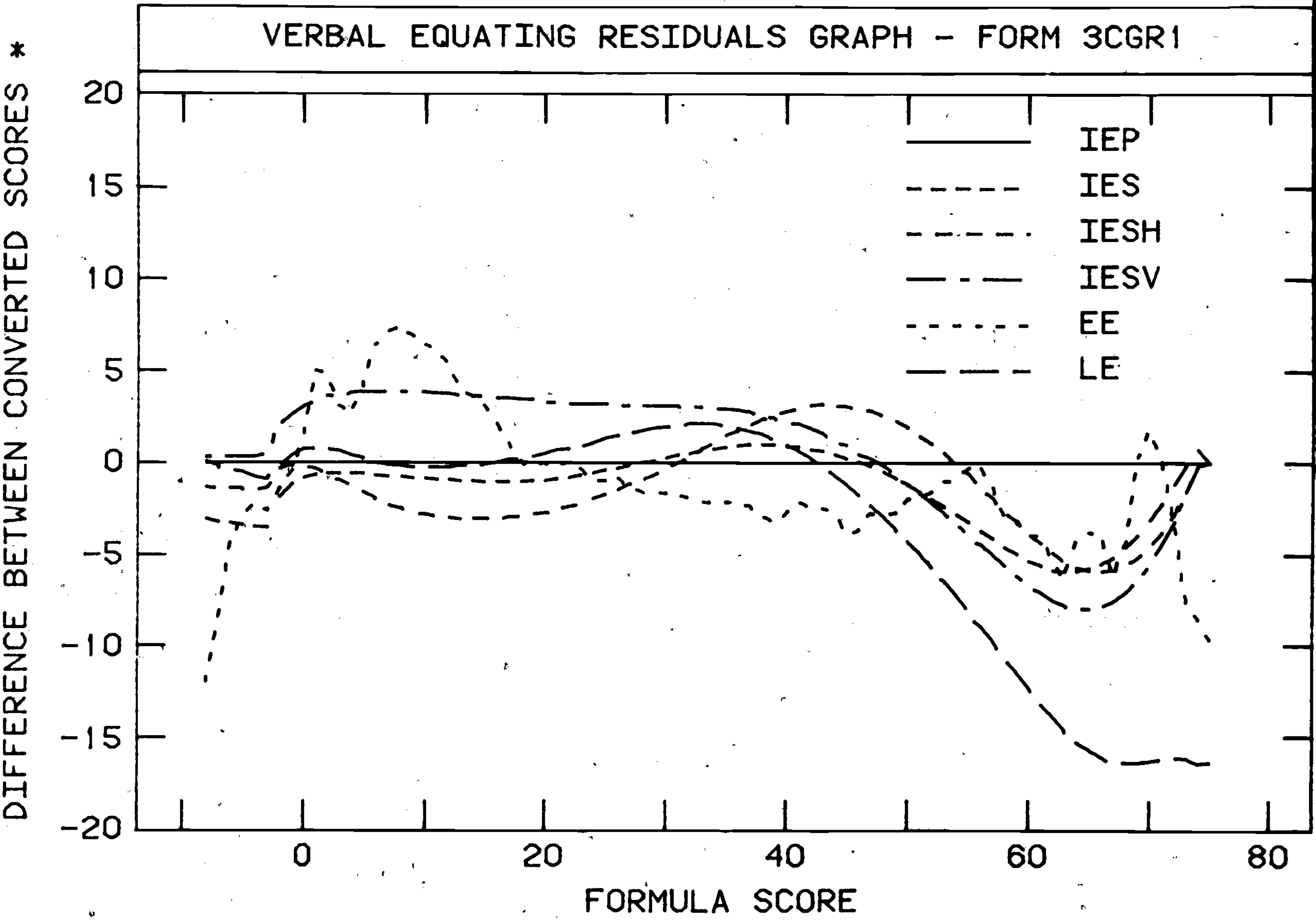
* SEE TEXT

Figure 22



* SEE TEXT

Figure 23



* SEE TEXT

Table 42

Quantitative Equatings
Means, Standard Deviations, and Skewnesses[a]

Equating Variant	Forms											
	3CGR1			K-ZGR3			K-ZGR2			ZGR1		
	Mean	S.D.	Skew	Mean	S.D.	Skew	Mean	S.D.	Skew	Mean	S.D.	Skew
IES	499.75	123.38	.15	*	*	*	*	*	*	*	*	*
IEP	494.81	123.65	.12	*	*	*	*	*	*	*	*	*
IAL	*	*	*	493.18	128.91	.04	530.09	127.48	-.11	526.55	133.75	-.10
IAL2	*	*	*	492.98	130.75	.04	*	*	*	*	*	*
EE	498.65	130.39	.01	*	*	*	*	*	*	*	*	*
LE	498.63	130.31	.17	486.06	134.94	.18	*	*	*	*	*	*
LA	*	*	*	*	*	*	525.55	133.33	-.01	524.50	133.47	-.07

[a] The cells in this table in which asterisks appear represent equatings that were not carried out in this study.

Figure 24

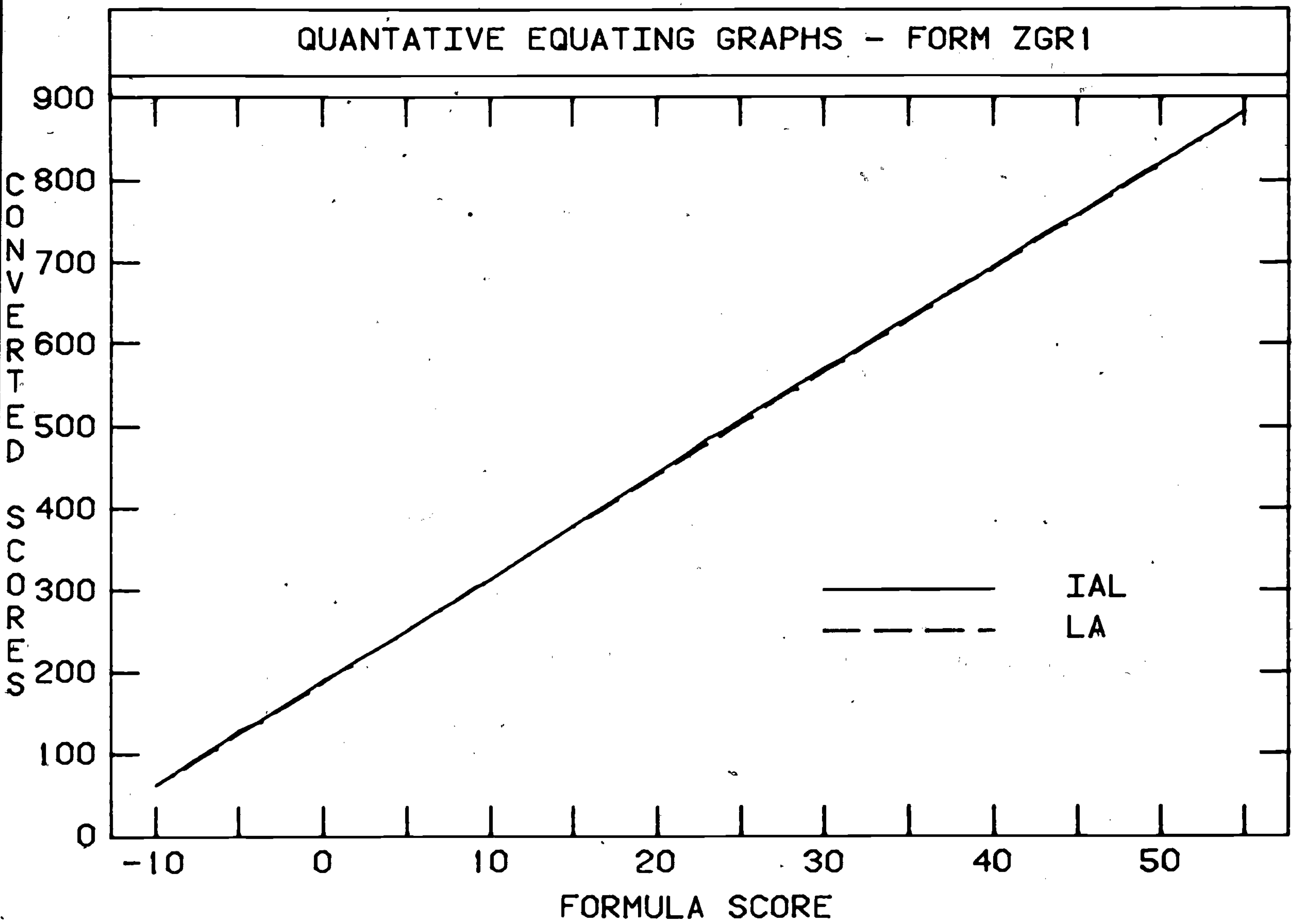

QUANTATIVE EQUATING GRAPHS - FORM ZGR1
CONVERTED SCORES
900
800
700
600
500
400
300
200
100
0
-10
0
10
20
30
40
50
FORMULA SCORE
IAL
LA

Figure 25

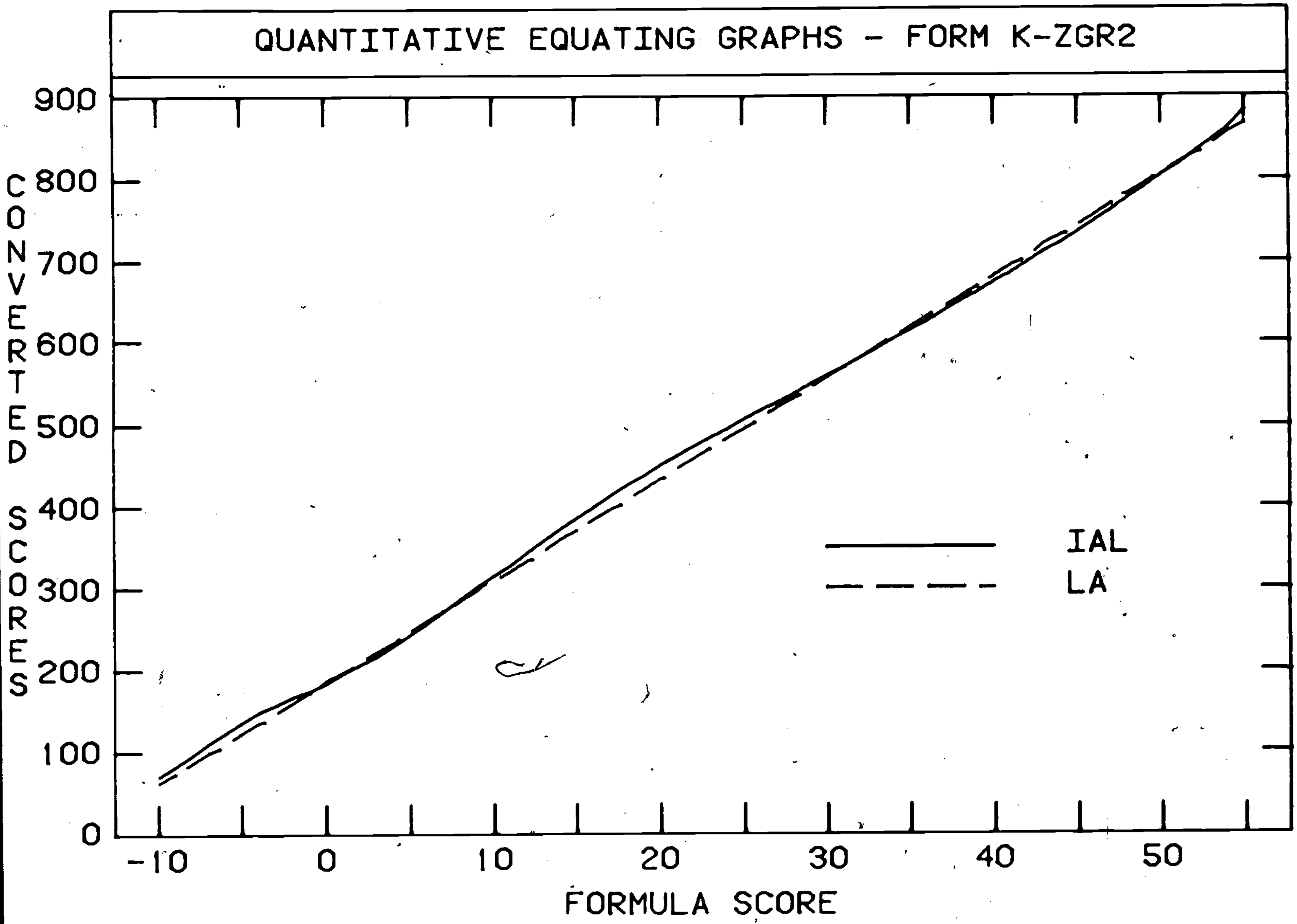

QUANTITATIVE EQUATING GRAPHS - FORM K-ZGR2
CONVERTED SCORES
900
800
700
600
500
400
300
200
100
0
-10
0
10
20
30
40
50
FORMULA SCORE
IAL
LA

Figure 26

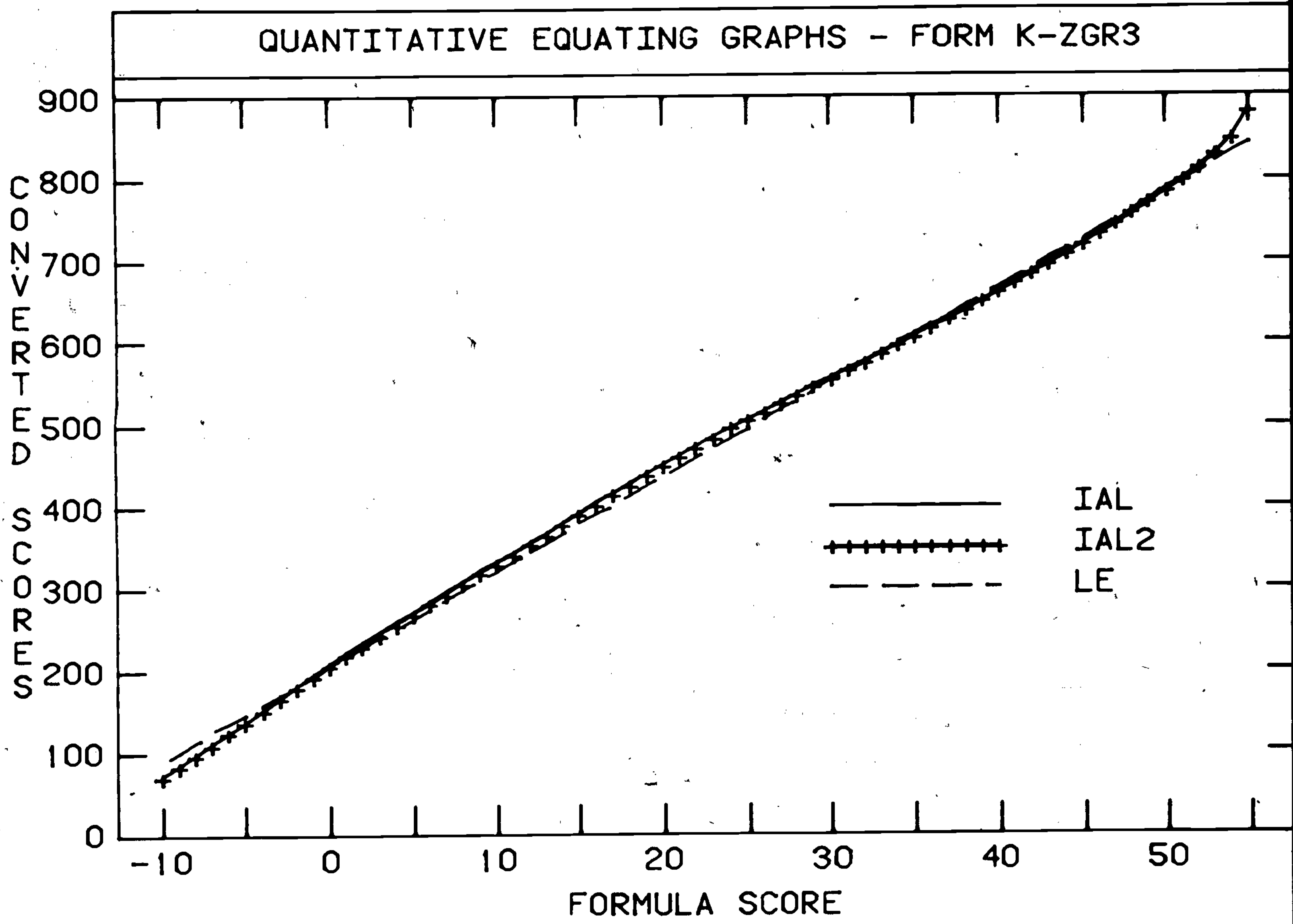

QUANTITATIVE EQUATING GRAPHS - FORM K-ZGR3
CONVERTED SCORES
900
800
700
600
500
400
300
200
100
0
-10
0
10
20
30
40
50
FORMULA SCORE
IAL
IAL2
LE

Figure 27

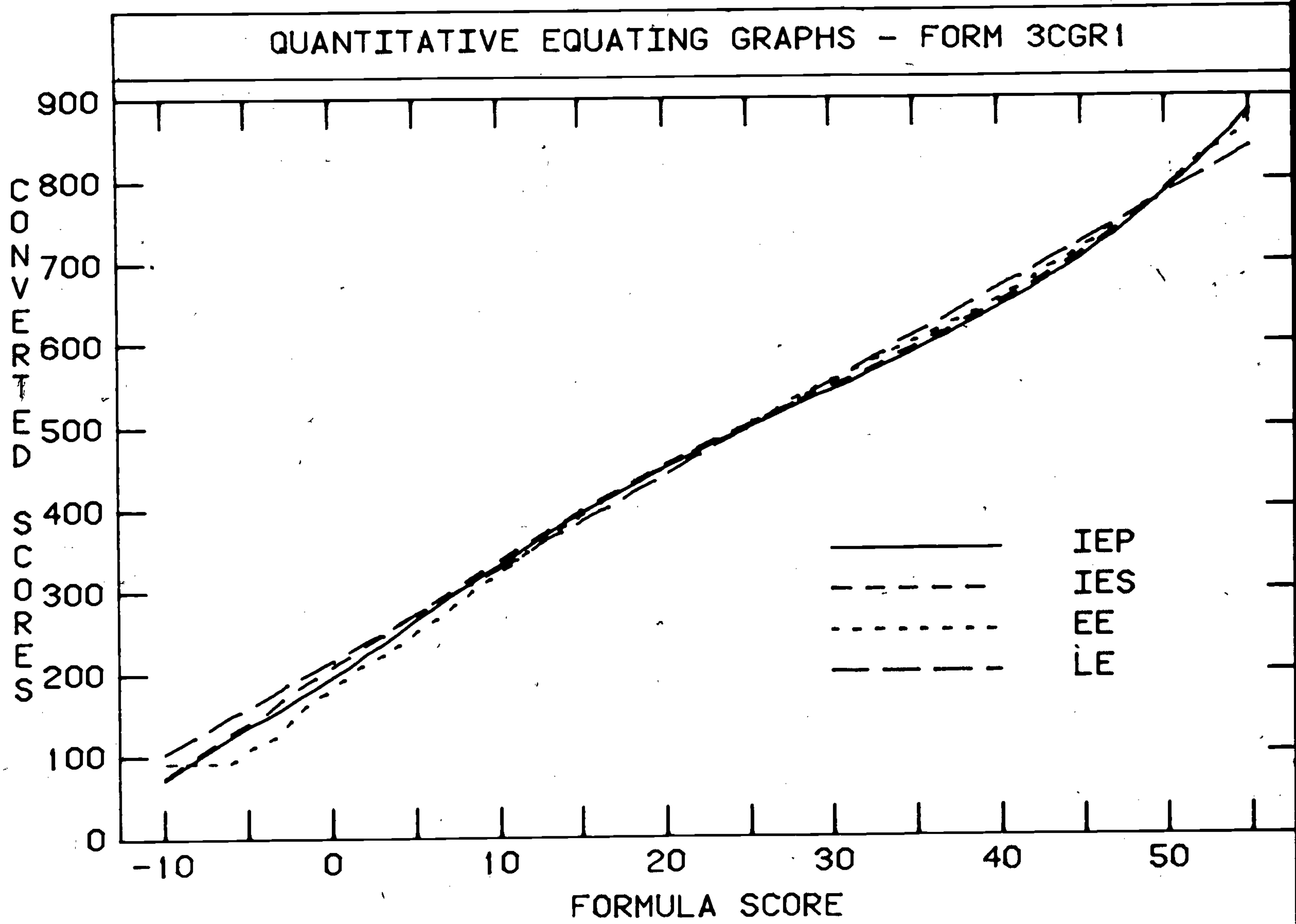

QUANTITATIVE EQUATING GRAPHS - FORM 3CGR1
CONVERTED SCORES
900
800
700
600
500
400
300
200
100
0
-10
0
10
20
30
40
50
FORMULA SCORE
IEP
IES
EE
LE

Figure 28

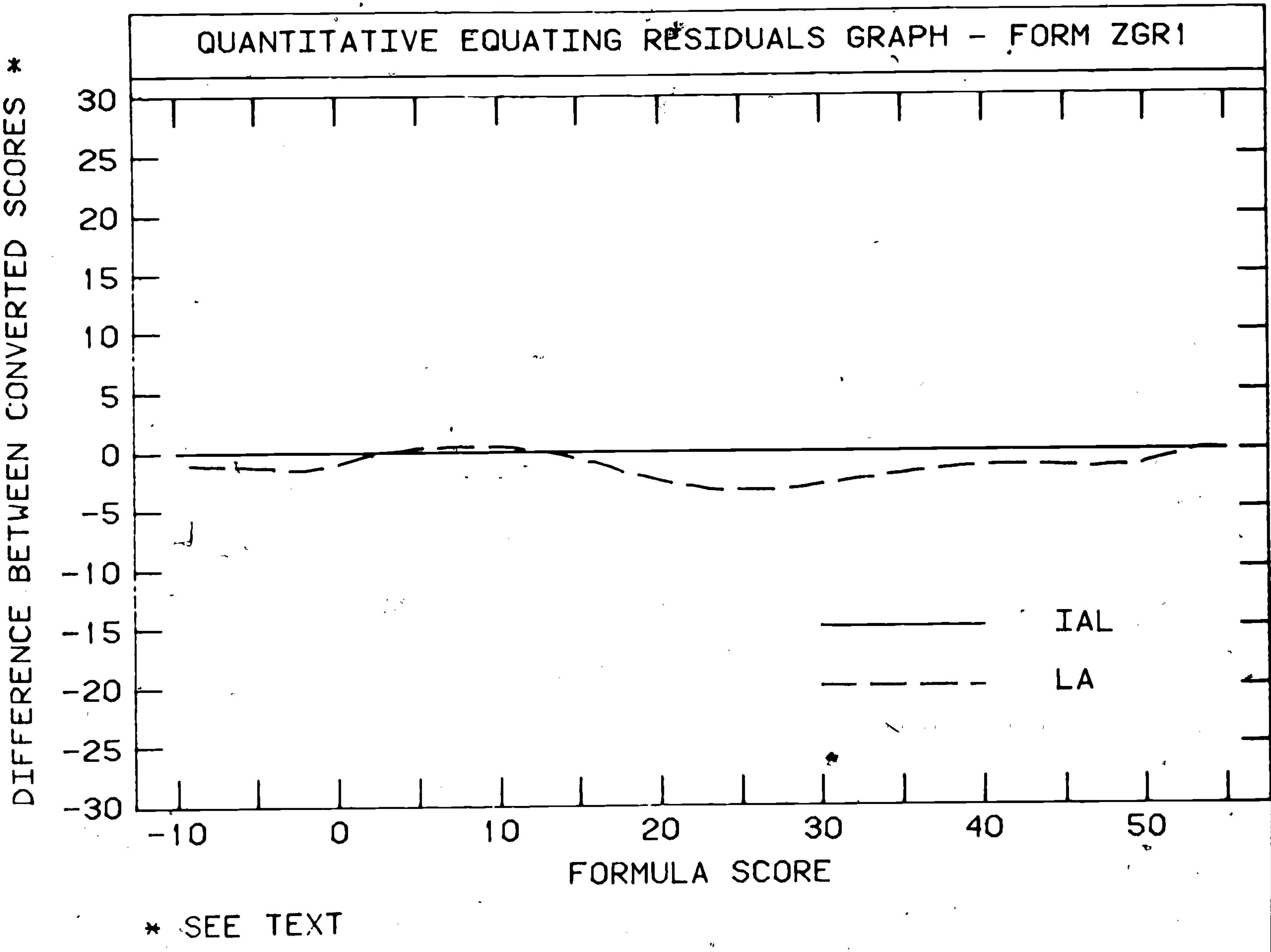

QUANTITATIVE EQUATING RESIDUALS GRAPH - FORM ZGR1
DIFFERENCE BETWEEN CONVERTED SCORES *
30
25
20
15
10
5
0
-5
-10
-15
-20
-25
-30
-10
0
10
20
30
40
50
FORMULA SCORE
IAL
LA
* SEE TEXT

Figure 29

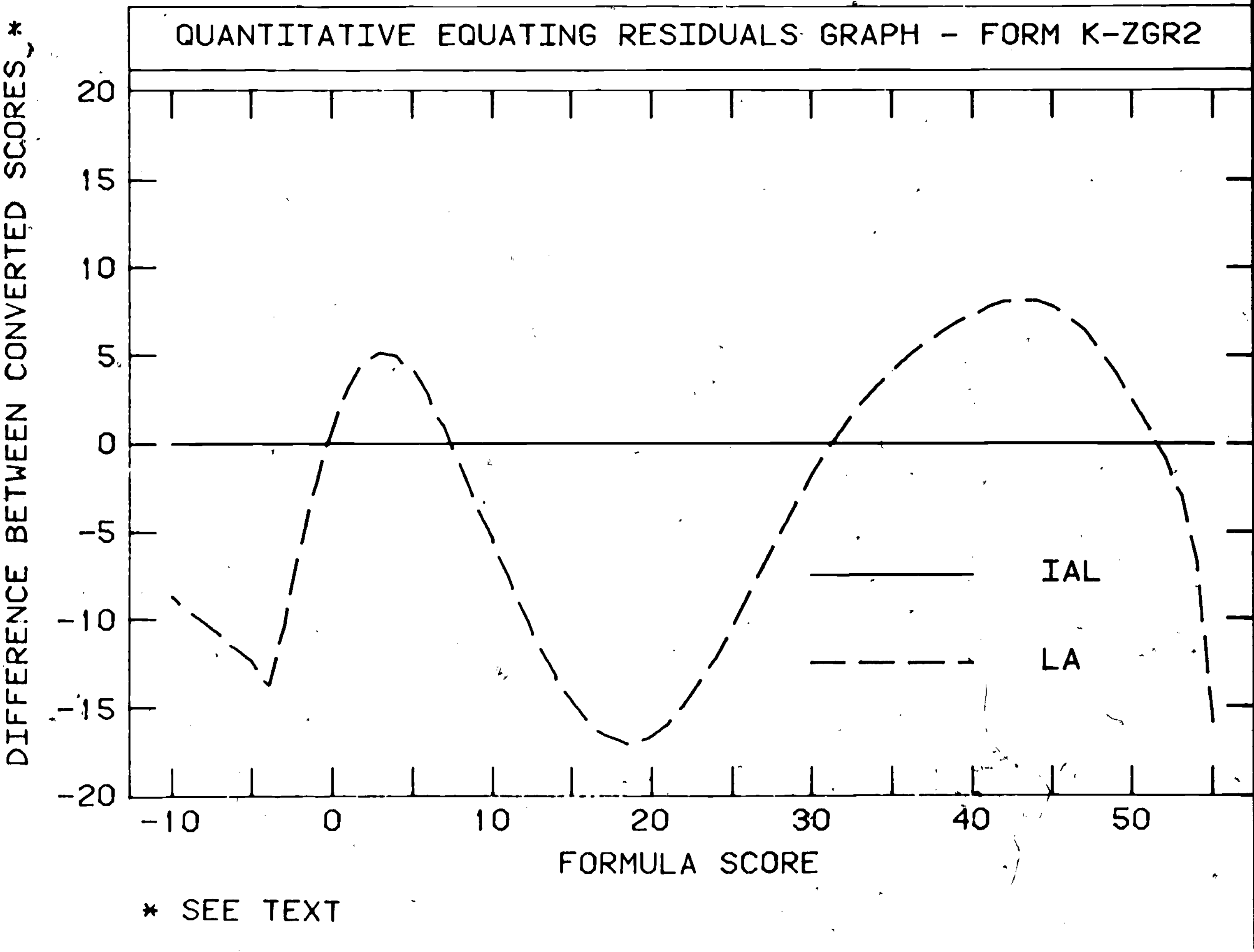
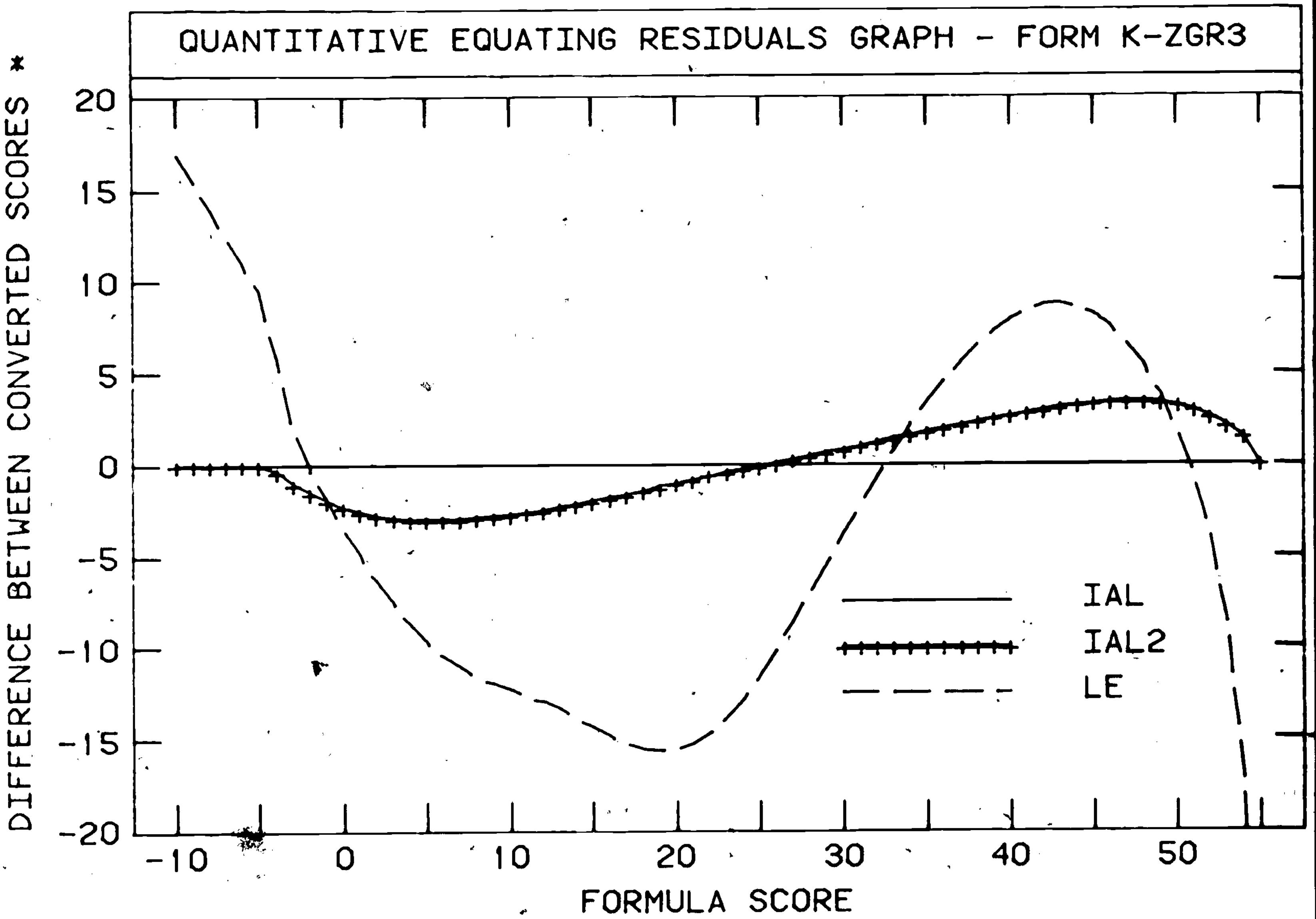
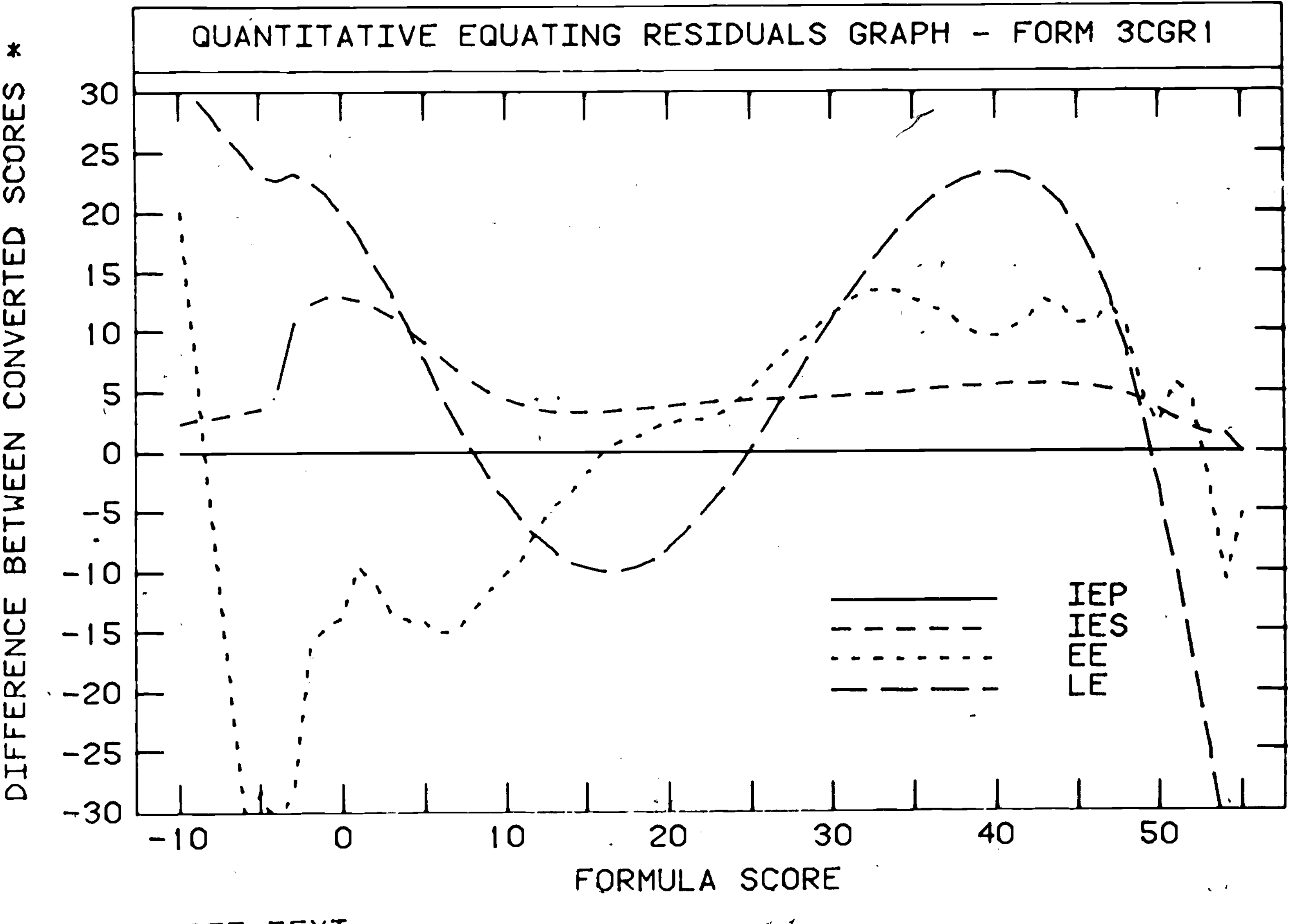

QUANTITATIVE EQUATING RESIDUALS GRAPH - FORM K-ZGR2
DIFFERENCE BETWEEN CONVERTED SCORES *
20
15
10
5
0
-5
-10
-15
-20
-10
0
10
20
30
40
50
FORMULA SCORE
IAL
LA
* SEE TEXT

Figure 30



* SEE TEXT

Figure 31



* SEE TEXT

Table 43

Analytical Equatings

Means, Standard Deviations, and Skewnessess

Equating Variant	Form 3CGR1		
	Mean	S.D.	Skew
IES	498.12	125.44	-.20
IEP	470.29	123.25	-.40
EE	497.37	128.95	-.29
LE	497.36	128.91	-.12

Figure 32

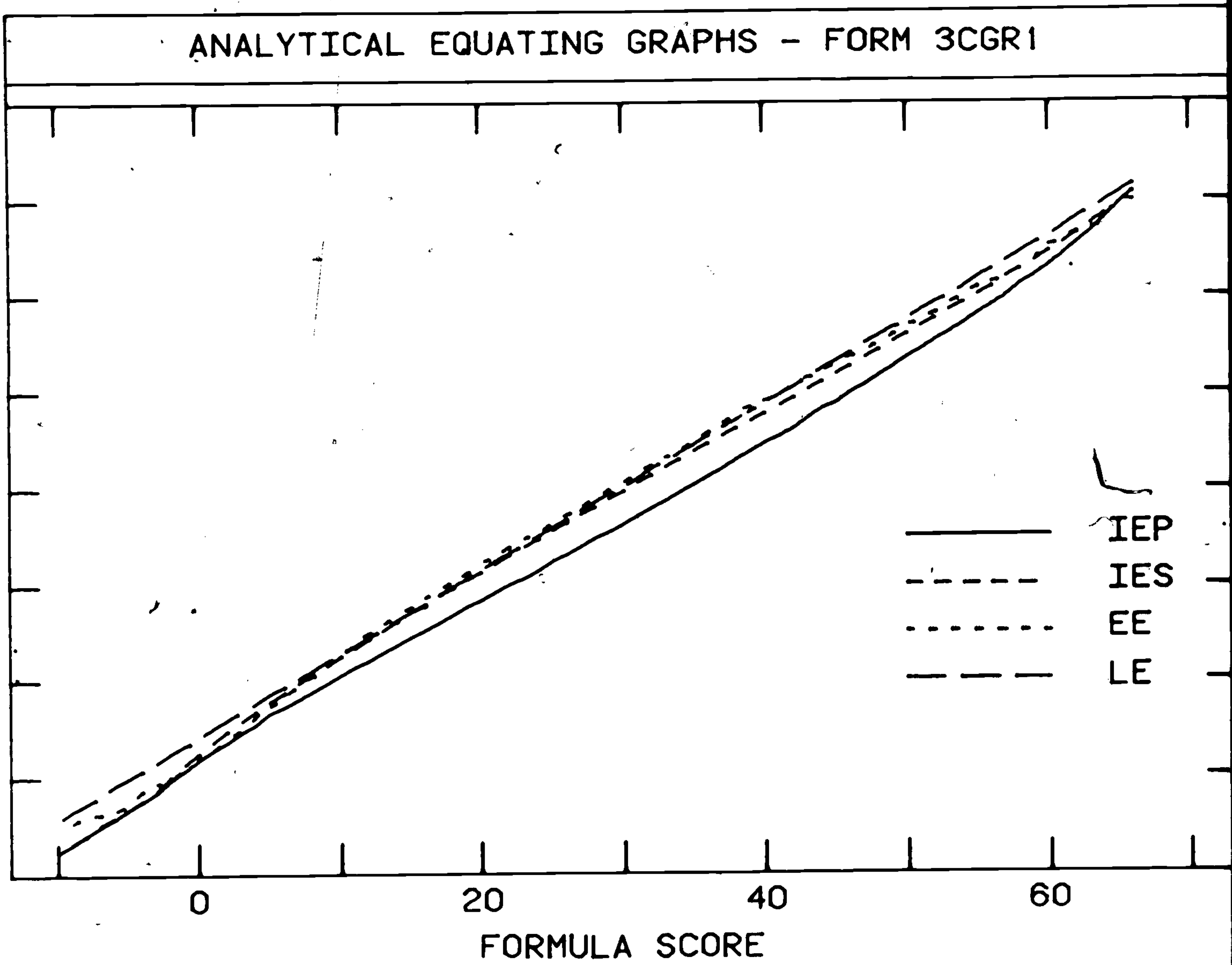
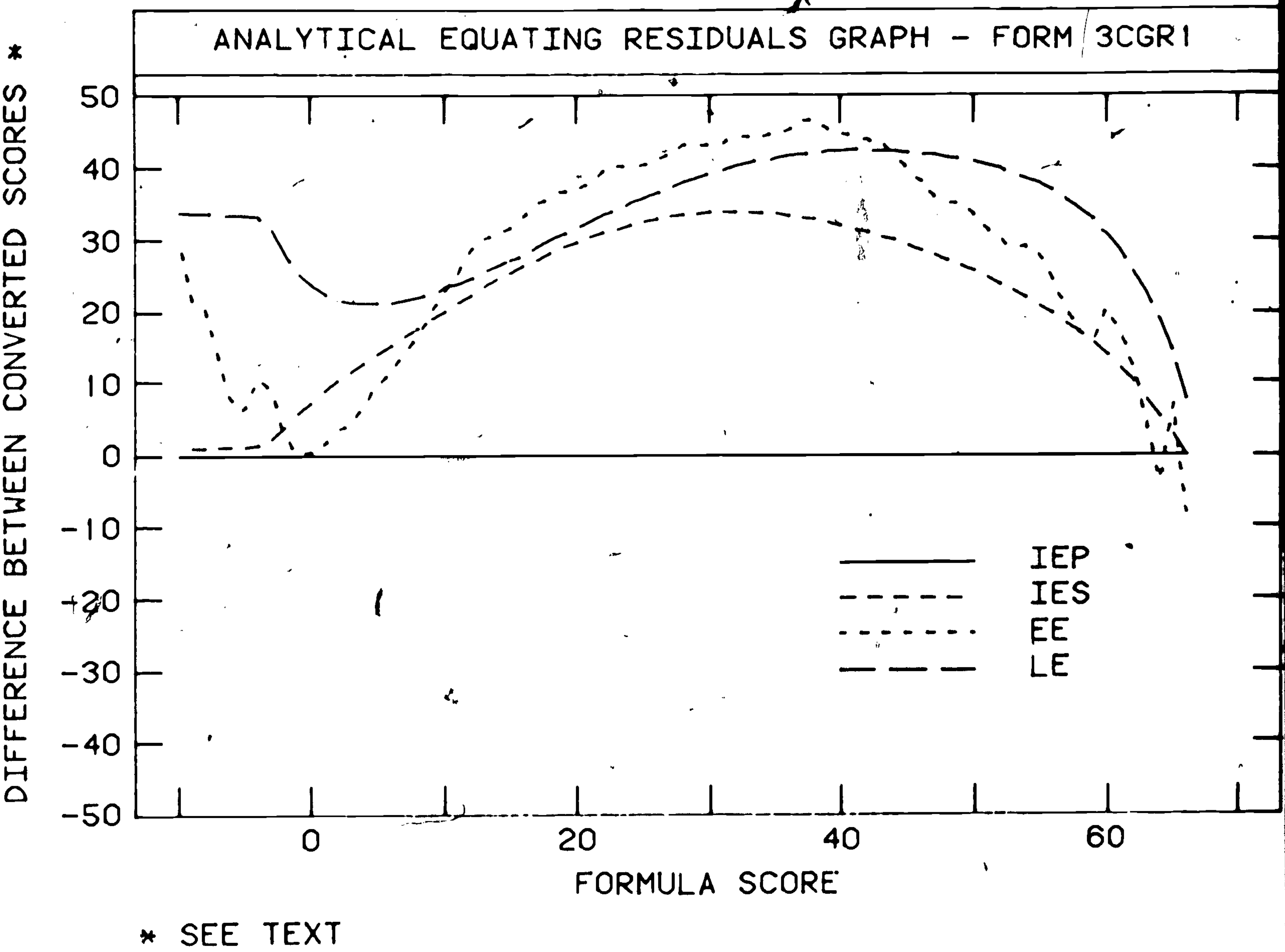

ANALYTICAL EQUATING GRAPHS - FORM 3CGR1
CONVERTED SCORES
900
800
700
600
500
400
300
200
100
0
20
40
60
FORMULA SCORE
IEP
IES
EE
LE

Figure 33



* SEE TEXT

is not possible and in the second case equating is not necessary. Assuming that we never have strictly parallel tests (and this assumption will be made throughout the rest of this chapter), and given the impossibility of equating fallible tests, one can still attempt to adjust scores as equitably as possible. The various equating models examined as part of this research are based on a variety of assumptions and are affected by a variety of factors. In order to judge the operational feasibility of IRT equating it is important to consider these factors and their potential though unknown effects on IRT, linear, and equipercentile equating methods and the equivalent-group and anchor-test data collection designs.

All equating, as mentioned previously, requires perfectly reliable tests. Additionally, all equating methods require that the tests to be equated are unidimensional (Morris, in press). How then do other assumptions (and the potential effect of violation of these assumptions) differ for IRT, equipercentile, and linear equating models?

Violation of the assumption of unidimensionality might lead to more serious consequences for IRT equating than for linear or equipercentile. This is because IRT is a stronger, more specific model; that is, IRT assumes unidimensionality explicitly at the item level. In contrast, all that is required for linear and equipercentile equating is unidimensionality at the test level. Each, however, requires unidimensionality in order to establish a single unambiguous, constant metric. Thus, the possible difference in effect of the violation of unidimensionality is unclear.

Some equating problems are based on the constraints of available data. The sparseness of data for low ability examinees makes it difficult to estimate the pseudoguessing parameter. Lack of appropriate data can also make it difficult to estimate the discrimination and difficulty parameters of very easy or very difficult items. Additionally, items that discriminate very poorly have poorly determined difficulty parameters (Kingston and Dorans, 1981, give an example of an item with parameters estimated on two samples of over 1,500 examinees where the estimate of b varied from more than +1.5 to less than -1.5). Similarly, equipercentile equating frequently suffers from a sparseness of data at the extremes of the score scale, which can lead to poor equating at those extremes. To a lesser extent, linear equating can be affected by outlying values having an undue influence on the mean and standard deviation. With the sample sizes typically used in equating the GRE Aptitude Test, however, this does not cause any difficulties.

Though Lord has shown that equating nonparallel tests requires perfect reliability, different equatings are probably differently affected by both imperfect reliability and differences in reliability between old and new test forms. It is likely that equating methods based on true score estimates (whether they are based on IRT or classical methods) are less adversely affected, at least by differences in reliability.

Even if a lack of parallelism between test forms is attributable solely to differences in item difficulties and/or discriminations (and is unrelated to multidimensionality), different equating methods will be influenced differently. Lack of statistical parallelism between forms results in a curvilinear equating relationship. We know that we cannot produce strictly parallel tests and that, if we could, equating would be unnecessary (Lord, 1980a). Thus, it is clear that linear equating can never precisely define the relationship between test scores on different forms. In many circumstances the departures from linearity appear minor, but, as test forms become less parallel, linear equating becomes less appropriate. Jaeger (1981) presents some experimental indices for investigating whether linear or equipercentile methods are more appropriate for equating.

Just as sparseness of data at the extremes presents a practical problem for some equating methods, discreteness of data can present estimation problems (Braun & Holland, in press; Potthoff, in press). Morris (in press) suggested linear equating might be preferable to equipercentile equating if there are "too few" items, but did not define "too few." Potthoff (in press) suggested not rounding formula scores before equipercentile equating or using IRT based equating to avoid problems caused by data discreteness.

Data collection designs necessitated by administrative complexities can lead to other problems with equating. The anchor-test design allows one to adjust for differences in examinee ability. There is evidence, however, that as the difference in ability between the two groups becomes larger, the quality of the equating based on the anchor design decreases (Marco, et al., 1979; Petersen, et al., in press). Since IRT equating is based on item parameters that are invariant with respect to examinee ability, it may be more resistant to this problem. This is supported by the Marco, et al. results.

The equivalent-group design, as it is typically used, based on practical considerations, also presents a problem. When an old and new form are spiralled, the old form has previously been exposed. Some of the examinees may have previously taken the old form and thus might be expected to perform better than their fellow examinees who have either taken a different old form or have not previously taken the test. Examinees taking the new form cannot experience a comparable benefit. Thus mean scores, to some small extent, may be artificially high on the old form compared to the new form and might consistently make the old form seem easier (although probably to an unnoticeably small extent) than it is. Such a systematic bias might lead to an eventual scale drift.

IRT based equating, as we have chosen to implement it, is not affected by speededness in the same way as are linear and equipercentile equating. To minimize multidimensionality, contiguous items to which an examinee has made no response and which appear at the end of a separately timed section were coded as "not-reached" and were not used in the estimation of the examinee's ability. Likewise, these "not-reached" items were not used in estimating the parameters of the items. Thus, the IRT method attempts to equate a more unidimensional ability metric. Since equating (as commonly used) provides a scaled score that is a function of an observed score, and since these observed scores have variance due to speededness, IRT equating based on item data including "not-reached" items might be subject to some problems that do not affect classical equating methods. If two forms of a test differ in speededness, IRT based equating, inappropriately, will not reflect this. The resulting bias in equating should be trivial if the variance due to the speed factor is very small compared to the variance due to the power factor or if the difference is speededness is quite small.

Verbal equatings. Table 41 shows that most verbal equatings produced similar results. Several findings are notable. As mentioned earlier, separate calibrations of discrete verbal and reading comprehension item sets were performed to investigate dimensionality. For both the equivalent-group design (IES, IESV, IESH) and the anchor-test design (IAL, IALV, IALH) the effect of multidimensionality and item calibration design was further investigated with three equatings for each test form (see the equating method section of this chapter for greater detail). If the verbal section of the GRE was perfectly unidimensional or if IRT equating was highly robust to violations of unidimensionality, there would be no systematic differences among the three equatings; the only differences would be due to sampling error. If dimensionality is a factor, one would expect the IES and IESV (or IAL and IALV) equatings to be more different from each other than from the IESH (or IALH) equating. Examination of Table 41 shows this to be the case. Surprisingly, there is very little difference between the IES and IESH equatings and IAL and IALH equatings. The difference between means based on the two equatings for forms 3CGR1, K-ZGR3, and ZGR1 are .12, .39, and .45, respectively. The difference between standard deviations is somewhat larger for one of the three forms: 1.37 versus .39 and .28. Form K-ZGR2 shows a somewhat larger discrepancy: 1.98 for the means and 1.84 for the standard deviations.

Form ZGR1 allows the most straightforward assessment of IAL, IALV, and IALH equating. In this one case, the LA equating is a true criterion since form ZGR1 has been equated to itself, and the LA statistics are based on given (and, for our purposes, we can assume arbitrary) scaling parameters that are also part of the IAL, IALV, and IALH equatings. The IALH equating is in closest agreement with the LA "scaling." This might simply be due to differential sampling fluctuations in the item parameter estimation procedure (almost but not quite identical samples were used in the three calibrations, see pages 18 through 31) but the possible superiority of the equatings based on homogeneous subsets deserves further investigation.

The IEP equating has summary statistics quite similar to the IES equating and not quite as similar to the LE and EE equatings. The IEP equating is based on a stronger parameter linking than the IES equating (spiralling versus a single LOGIST run), but the IEP estimates are potentially subject to a practice or item position effect. Kingston and Dorans (1982) have shown that the position of GRE verbal items when administered has no systematic effect on item parameter estimates. Several factors could be responsible for the differences (though relatively small) between the IES and IEP and the LE and EE equating results. Though the relative efficiency graphs (see Appendix B, figures B.4a through B.4d) do not show evidence of a lack of parallelism, form ZGR1 was more speeded than form 3CGR1 (80 percent of the examinees taking a spiralled subform (C47) of ZGR1 reached 61 items; 80 percent of the examinees taking a similar subform (C41) of 3CGR1 reached 65 items). Unlike the equatings for forms K-ZGR3 and K-ZGR2, the equatings (IRT, linear, and equipercentile) for form 3CGR1 were all based on samples from the same data (EE and LE were based on identical data; IES, IESV, and IESH were all based on an almost identical subset of the EE, LE data; IEP was based on an essentially random one half of the IES sample).

The IAL and IAL2 equatings of form K-ZGR3 have very similar summary statistics. The minor differences (.71 between means, .06 between standard deviations) are a result of the extra link in the parameter scaling using the IAL2 method. It is encouraging to see that these differences are small. Ignoring the IALV and IAL2 equatings since there is no theoretical reason for ever prefering them, the means and standard deviations for the IRT equatings (IAL, IALH, IAW) are more similar to each other than they are to those of the LE and EE equatings. Much of this difference might be attributable to differences in the groups on which the equating data are based. The three IRT equatings were based on data from a different group of examinees than that available for the LE and EE equatings. It should be noted that the .95 confidence interval of the LE equating is no smaller than +2.16 scaled score points at its smallest point, the mean of the distribution (based on data given in Stewart, 1981).

The results of the K-ZGR2 verbal equatings are less clearcut. The means based on each method differ from all other means by at least 1.02 and range from 496.68 (IAL) to 503.18 (IAW). The standard deviations (ignoring IALV) range from 123.30 (IALH) to 126.26 (LA). The two most similar results are for IAW and LA (difference in means was 1.02, difference in standard deviations was .60).

Quantitative equatings. The quantitative equatings, as compared using the means and standard deviations given in Table 42, appear to be less similar than the verbal equatings. For form ZGR1, the linear equating parameters from which the LA data were derived are part of the scaling for the IAL equating. Thus, we would expect to reproduce the LA mean and standard deviation quite closely. The difference in standard devations (.28) is acceptably small. The difference in means appears somewhat large (2.05). Unfortunately, we do not have an estimate for the standard error of equating for the IAL method to help put these differences in perspective.

All four quantitative equatings performed on form 3CGR1 were based on data from the same administration. The IES mean was not so different from the EE and LE means (1.12), but the IES and IEP means differed by 4.94 scaled score points. Even more striking is the difference between the IRT based standard deviations and the EE and LE based standard deviations, approximately 7 scaled-score points. While the parameter estimates for the 3CGR1 items were based on samples of only about 2,500 for the IEP equating and about 5,000 examinees for the IES equating, it seems unlikely that these differences can be attributed solely to sampling fluctuation in the parameter estimation process. Although the difference in means between IEP and IES is in the direction that would be expected if there were a practice effect (items being easier when calibrated in the fifth section than when calibrated in the operational section), Kingston and Dorans (1981) investigated practice effect on the item level and found no evidence supporting this hypothesis for quantitative items.

Figure 27 compares the equating lines for the various methods used on 3CGR1 quantitative. The most striking result is the marked curvilinearity of the IRT equatings. The EE equating is also quite nonlinear, although not as much as the IRT equatings. The relative efficiency curves provide direct evidence of marked nonparallelism of these two forms (Appendix B, Figures B.8a and B.8b). In addition, examination of the formula raw score data for spiralled samples based on subforms C49 (ZGR1) and C44 (3CGR1) provides evidence of differential speededness. On ZGR1, 80 percent of the examinees reached item 50, on 3CGR1, 80 percent of the examinees reached item 48. Similarly, on ZGR1, only 50.1 percent of the examinees completed the test while, on 3CGR1, only 34.8 percent finished the test. These results must be considered in light of the difficulty of the two forms. The mean raw score of the ZGR1 sample was 24.59, in the 3CGR1 sample, it was 24.52. Thus, since the forms contained the same number of items, the forms are of different speededness, and this might bias the IES and IEP equatings.

Results for the K-ZGR3 and K-ZGR2 equatings are also difficult to interpret. The means and standard deviations based on the IRT equatings differ from the results of the linear equatings. For the IRT equatings, we know there are potential problems with dimensionality and model fit. For the K-ZGR3 quantitative equating, LE is really a complex combination of equatings. The base of that series was the equating of the original K-ZGR3 to ZGR1. Figure B.7a provides evidence that these forms have markedly nonparallel quantitative sections. This explains the curvilinearity of the IRT equatings for K-ZGR3, and the consistency of this nonparallelism for quantitative forms suggests that the appropriateness of linear equating for the quantitative section of the GRE should be further investigated.

Analytical equatings. Statistics based on the analytical equatings of form 3CGR1 are presented in Table 43. The most noticeable result is the extremely low mean based on the IEP equating. This difference of 27.07 points between the IEP mean and the LE (the least different) mean is due to practice effect, most noticeably on the analysis of explanations items. This effect is more fully documented by Kingston and Dorans (1981).

The mean and standard deviation for the IES equating are somewhat different from those for LE and EE (.75 and 3.51 between IES and EE). The relative efficiency graph (B.9a) and the curvilinearity of both the IES and EE equatings suggest that the LE equating is not appropriate because of the nonparallelism of the two forms. Problems with the model fit of analysis of explanations items and the complex factor structure of the analytical section further complicate the interpretation of these results.

Shifts in dimensionality. A general consideration for interpreting the results of GRE equatings is the possibility of shifts in the dimensional characteristics of the test sections due to nonrandom choice of administration dates by markedly different types of students. Mathematics and science oriented students tend to take the GRE Aptitude Test in the fall while social science and education students tend to take the test in the spring. It is likely, to the extent that this difference in factor structures across administrations exists, that all equating methods will be somewhat affected, although perhaps to different degrees.

SUMMARY, DISCUSSION, AND RECOMMENDATIONS

The research reported here is based on tne GRE Aptitude Test as it was structured during the period from December 1979 through June 1980. At an early stage of this research it was decided that the analytical section would soon undergo substantial revision. Consequently, this research focuses on the verbal and quantitative sections. Moreover, in October 1981 the verbal and quantitative sections and the general structure of the entire GRE Aptitude Test were revised. Factors from this restructuring that are most likely to affect the use of item response theory are the increase in the time-per-item allowance, changes in the relative proportions of certain item types, and the shift from formula to rights only scoring. It is difficult to forecast tne exact effects of these changes. Recommendations to be presented will be influenced by expectations about the effects of these changes.

This final section of the report summarizes the findings of the various portions of the research, and then synthesizes these findings. The topics to be summarized are: the basic assumptions of item response theory, implications of previous factor analytic research conducted on the GRE Aptitude Test, assessment of the weak form of local independence, analysis of item-ability regressions, temporal stability of item parameter estimates, sensitivity of parameter estimates to violations of unidimensionality, and comparisons of item response theory equating with equipercentile and operational linear equating.

Summary

The basic assumptions of item response theory. One of the major assumptions of item response theory is that performance on a set of items is unidimensional, i.e., the probability of successful performance by examinees on a set of items can be modeled by a mathematical model with only one ability parameter. A second major assumption is that tne probability of successful performance on an item can be adequately described by the three-parameter logistic model, a particular item response theory model that seems particularly applicable to binary-scored multiple-cnoice items.

One consequence of the unidimensionality assumption is the mathematical concept of local independence. The weak form of local independence, which was assessed in this research, states that item responses are uncorrelated at fixed levels of ability, i.e., after taking ability into account, tnere are no systematic shared influences on item performance.

Implications of previous factor analytic research on the GRE Aptitude Test. Four factor analytic research studies conducted on the GRE Aptitude Test were reviewed in order to assess the dimensionality of the test, to identify sets of homogeneous items, and to extract hypotheses about the GRE Aptitude Test that could be tested in other phases of this research. The four factor analytic studies provided strong evidence for the existence of three large global factors: general quantitative ability, reading comprehension or general verbal reasoning ability, and vocabulary or discrete verbal ability. In addition, the factor analytic studies provided evidence for the existence of several smaller factors: a data interpretation factor, a technical reading comprehension factor, and a speed factor on the verbal scale.

As a consequence of these studies, verbal items were separated into a reading comprehension set and a discrete verbal set for the purposes of item response theory analyses. However, the studies also suggested separation of the data interpretation items from other quantitative items and the further breakdown of reading comprehension items into a set of technical reading comprehension items and a set of other reading comprehension items. Doubts about the practical significance of these smaller dimensions, coupled with the fact that there were too few items to yield stable linking of ability scales through item response theory item-difficulty estimates, led to the conclusion that the construction of separate data interpretation and reading comprehension scales was not feasible, given the current structure of the GRE Aptitude Test.

Assessment of the weak form of local independence. The weak form of local independence states that, for a given ability level, item responses are uncorrelated. This local independence condition was assessed via the examination of item intercorrelations with estimated ability partialled out. Partial correlations both with and without a correction for guessing were examined.

The analysis of partial correlations for the verbal subtest uncovered two systematic sources of local independence violations: a reading comprehension factor and speededness. The analysis of partial correlations for the quantitative test revealed that the data interpretation items retained positive intercorrelations after overall quantitative ability was partialled out, thus providing evidence for another source of local independence violations. In sum, the partial correlation analyses produced findings consistent with expectations based on the previous factor analytic studies.

Analysis of item-ability regressions. The item response function of item response theory can be viewed as a theoretical form for the regression of item score (1 = a correct response, 0 = an incorrect response) onto underlying ability. Actual item performance for each ability level can be obtained from the data and plotted for various levels of ability to obtain an empirical item-ability regression. Comparisons of estimated item response functions to actual item-ability regressions enable one to assess the fit of the three-parameter logistic model to the data. A graphical technique, referred to as analysis of item-ability regressions, was devised to assess fit via these comparisons of estimated and empirical item-ability regressions.

On the basis of the analysis of item-ability regressions, it was determined that all of the verbal item types and two of the analytical item types, logical diagrams and analytical reasoning, seemed to be fit better by the three-parameter logistic model than the three quantitative item types and the analytical analysis of explanations item type. Of these latter four item types, regular mathematics and data interpretation items seemed to be fit only a little less well than some of the better-fitted item types. Quantitative items were the most difficult items for the three-parameter logistic model to fit. Analysis of explanations items keyed other than B or E were fit by the model quite well, but those keyed B or E had the highest proportion of model fit scores that indicate poorer fit of any of the item classifications under study.

Temporal stability of item parameter estimates. Theoretically, an item response function for an item should not be affected by when the item was administered, provided a common ability metric has been established. The section on parameter estimation and item linking procedures described the procedures used to place all item parameter estimates on the same scale. The dual administrations of Form ZGR1, once in February 1980 and once in June 1980, enabled us to assess the temporal stability of item parameter estimates.

For the discrete verbal items, the item difficulty parameter, b_g, the item discrimination parameter, a_g, and the item response function derived estimate of conventional item difficulty, p_g, all exhibited much temporal stability. The psuedoguessing parameter, which is the most difficult parameter to estimate, exhibited less temporal stability.

For the reading comprehension items, b_g, a_g and p_g all exhibited much temporal stability. The c_g estimates, however, were much more sensitive to administration date.

All quantitative items had very stable item parameter (a_g, b_g, and c_g) estimates, and very similar conventional item difficulty estimates, p_g, over time.

Sensitivity of parameter estimates to violations of unidimensionality. Evidence indicating that verbal items are not homogeneous, i.e., that they measure more than one dimension, was presented in the factor analytic review, the assessment of local independence, and the item-ability regressions. Comparisons of item parameter estimates based on calibration of heterogeneous sets (all verbal items) and homogeneous sets (discrete verbal only or reading comprehension only) were suggested by these earlier results.

Discrete verbal and all verbal calibrations of discrete verbal items produced considerably more simiiliar estimates of item discrimination than the reading comprehension and all verbal calibrations of reading comprehension items. The discrete verbal and all verbal calibrations produced slightly more simiiliar estimates of item difficulty, b_g, for the discrete verbal item than the reading comprehension item estimates of b_g produced by the reading comprehension and all verbal calibrations. When compared to the results for a_g estimates, the b_g estimates exhibited much less sensitivity to homogeneity of item sets.

With the exception of the c_g estimates of the discrete verbal items of form K-ZGR2, the c_g estimates appeared fairly robust to heterogeneity of item calibration set. The exceptional results obtained for the discrete verbal items of form K-ZGR2 were an artifact produced by the choice of constraints used by LOGIST to estimate c_g for items that are deemed too easy to provide well-determined estimates of c_g. Compared to a_g and b_g estimates, however, the c_g estimates reflected greater sensitivity to item heterogeneity, a result partly reflecting difficulties inherent in obtaining stable estimates of c_g.

The similarity of p_g estimates based on heterogeneous versus homogeneous calibrations was very high. An inference suggested by this high degree of similarity is that the observed data can be approximated equally as well by sets of heterogeneous items (all verbal) as by sets of homogeneous items (discrete verbal, or reading comprehension).

Comparability of ability estimates based on homogeneous and heterogeneous sets of items. All verbal items were calibrated at least twice, once with a set of homogeneous items of like type, e.g., discrete verbal or reading comprehension, and once with a set of heterogeneous items comprised of both discrete verbal and reading comprehension items. This procedure produced three ability scores for each examinee: a verbal ability score based on all verbal items, a discrete verbal ability score based on discrete verbal items, and a reading comprehension score based on reading comprehension items. Correlations among these ability estimates and among proportion-correct true scores based on these ability estimates provided evidence for the existence of two distinct, highly correlated reading comprehension and discrete verbal abilities. Evidence was also provided for thinking of the overall verbal ability score as a weighted composite of the discrete verbal and reading comprehension abilities. Although the overall verbal ability score appears to have resulted from LOGIST being drawn toward the discrete verbal dimension during parameter estimation iterations, the correlations it has with the discrete verbal and reading comprehension abilities are consistent with the correlations one would expect if the overall verbal proportion-correct true score were defined as a weighted composite of the discrete verbal and reading comprehension true scores, where the weights were relative number of discrete verbal and reading comprehension items, respectively. Of course, the correlation between discrete verbal and reading comprehension abilities is high enough to ensure that any set of positive weighting coefficients would produce a composite dimension that was proximate to the verbal dimension. In sum, the evidence provided support for the existence of two distinct, highly correlated discrete verbal and reading comprehension abilities that can be combined to produce a composite ability that closely resembles the general verbal ability dimension defined by LOGIST.

Equating comparisons. A statistical equating method is an empirical procedure for determining a transformation to be applied to the scores on one form to produce scores that are on the same scale as the other form. As such it consists of two parts, a data collection design and a set of rules for determining the transformation. Two data collection designs (equivalent group and anchor test) and three general statistical methods of equating (equipercentile equating, linear equating, and item response theory based true score equating) were used in this research.

In general IRT equating methods seemed to give reasonable results for the verbal equatings. The results for the quantitative section equatings are more questionable for several reasons: the relatively poor model fit of the quantitative items, particularly quantitative comparison items, and the possible shifts in dimensionality due to nonrandom choice of administration dates by markedly different types of students. That is, mathematics and science oriented students tend to take the GRE Aptitude Test in the fall and social science and education students tend to take the test in the spring. Results for the analytical section are marked by the large practice effect for IEP equating. The IES equating seems reasonable.

Synthesis

The major purposes of this research were to address the reasonableness of the assumptions of item response theory and the robustness of item response theory methods (applied to the GRE Aptitude Test) to violations of these assumptions. The research was motivated by a need to address the psychometric feasibility of applying IRT methods to the GRE Aptitude Test items and populations. Test disclosure legislation and its effects on operational equating strategies served as a major impetus for the need to address psychometric feasibility. If applicable to the GRE Aptitude Test, item response theory would provide powerful, flexible tools for in-depth analysis of test forms and items, the maintenance of score scales via equating, and the development of better and more efficient test forms that could be tailored to fit specific needs.

Fit of item response theory model to the GRE Aptitude Test items and examinee populations. Any evaluation of the fit of a mathematical model to data should be made from a realistic point of view that recognizes that all models are the products of human minds that attempt to understand and predict phenomena. As such, models never completely fit the data. Fit is a matter of degree.

The three-parameter logistic model seems to fit the GRE Aptitude Test data reasonably well for verbal and less well for quantitative and analytical. Evidence exists for the violation of local independence on all three scales of the test. On the verbal scale, the factors underlying reading comprehension items, particularly technical reading comprehension items, and speededness contribute to the lack of fit of the three-parameter logistic model to verbal items. Despite the existence of these sources of local independence violations, the model fits all verbal items reasonably well, as evidenced in the item-ability regression analysis, the relative insensitivity of item parameter estimates to homogeneity of item parameter estimation sets, and the verbal equating results. The shift to number right scoring will probably not enhance the fit of the three-parameter logistic model to verbal item types. The increased time per item should diminish discrepancies between IRT and other equatings when forms are differentially speeded.

On the quantitative scale, the data interpretation items were influenced by some systematic source of local independence violations, as evidenced in the chapters on the factor analysis review and the assessment of the weak form of local independence. The item-ability regression analyses and the equating results demonstrated that the three-parameter logistic model does not fit the quantitative items as well it fits the verbal items. The quantitative comparison item type was the most difficult item type to fit; there were some instances of marked nonmonotonicity of empirical item-ability regressions for this item type. The relative lack of statistical parallelism of the quantitative tests probably contributed to the greater dissimiliarity between scaled score distributions produced by the IRT methods and those produced by the operational linear method.

The three-parameter model fits the verbal items better than the quantitative items despite the fact that the dimensionality analyses appear to indicate that dimensionality is a greater problem with the verbal item types than with the quantitative item types.

Application of the common factor model, a linear model, to the GRE Aptitude Test, clearly identified two major verbal dimensions, reading comprehension and discrete verbal, as well as some minor dimensions. On the other hand, factor analyses of the quantitative items did not produce two clearly defined major dimensions. Perhaps, however, the subtle dimensionality problems implied by the item-ability regression analysis present a greater problem for the quantitative scale than does the grosser multidimensionality of the verbal scale. The verbal scale appears to be composed of two clearly defined, highly correlated dimensions that are amenable to modelling by a two-factor linear model. The high correlations between the two dimensions indicate that, while distinct, the two major categories of items are not very far from being considered functionally homogeneous. As a consequence of this functional homogeneity, the three-parameter logistic model fits the verbal data well, and the results of IRT and linear equating are to a large degree similiar.

In contrast, the quantitative scale does not seem to be fit as well by either the nonlinear three-parameter logistic model nor a linear model. As a consequence, the linear common factor model does not describe quantitative data as well as it does verbal data and is, therefore, less useful as a tool for accurately assessing the dimensionality of the quantitative items. In other words, the quantitative scale may be composed of heterogeneous items that are influenced by multiple dimensions that can not be adequately described by the linear common factor model. Empirical evidence for this hypothesis exists in the relative efficiency curves for the quantitative subtests and the observed correlations between the different quantitative item types. The former demonstrate a relative lack of statistical parallelism, while the latter demonstrate that data interpretation items share relatively little in common with other quantitative items.

The three-parameter logistic model does not fit analytical items as well as it fits verbal items. The soon-to-be-replaced analysis of explanations item type is the major source of local independence violations. This item type is very susceptible to practice effects, which are problematic for the precalibration (IEP) method of IRT equating. In addition, these items exhibit instances of nonmonotonic empirical item-ability regressions, when the keyed response is option B or E. Due to the planned major overhaul of the analytic section, this research did not focus on this section. The analytical section was examined closely enough to confirm the wisdom of the decision to remove the analysis of explanations item type. More complete evidence for the wisdom of this decision is contained in Kingston and Dorans, 1982.

Applicability of item response theory equating methods. The aspect of this research with the most direct bottom-line implications is the equating comparisons. Due to test disclosure legislation, the current linear method may no longer be a feasible equating procedure. A replacement or supplement should be found. Item response theory equating is particularly desirable because of other powerful statistical tools it provides in addition to equating. Lord (1980a) describes several of these powerful tools that item response theory can supply to the testing world. In this research, six different variants of item response theory true score equating were examined. Of these six approaches, the precalibration (IEP) method holds the most promise for coping with the constraints imposed by test disclosure legislation. Unfortunately, it is the IRT method most susceptible to practice effects, as witnessed in the analytical equatings of form 3CGR1. The other sections of the Aptitude Test do not show this practice effect, but a subtle effect that causes a systematic scale drift might exist. Consequently, the susceptibility of particular item types to practice effects determines, to a large extent, the feasibility of using the IEP method for equating.

While a companion report describes practice effect in detail, a summary of these findings suffices for our purposes of assessing the feasibility of using the IEP method of IRT equating on the GRE Aptitude Test. The discrete verbal item type is not susceptible to practice effects. The reading comprehension item type shows evidence of a possible fatigue effect. While the analysis of explanations items are very susceptible, neither logical diagrams nor analytical reasoning items are very susceptible. None of the quantitative item types appear to be susceptible to practice effects.

In sum, the item response theory model and the precalibration method of IRT equating are most applicable to verbal item types, less applicable to quantitative item types because of dimensionality problems with data interpretation items and instances of nonmonotonicity for quantitative comparisons items, and least applicable to the existing analytical item types because of the severe practice effects associated with the analysis of explanations item type and its other problems. Planned revisions of the analytical section, particularly the removal of the troublesome analysis of explanations item type, should enhance the fit and applicability of the three-parameter model to the analytical scale. Planned revisions to the verbal section are not expected to affect greatly the satisfactory fit of the model to verbal item types. It is unlikely that planned revisions will improve the appropriateness of IRT methods for the heterogeneous quantitative scale. A fuller understanding of the workings of this rather complex scale is needed.

REFERENCES

Bejar, I. A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 1980, 17, 283-296.

Braun, H. & Holland, P. Observed score test equating: a mathematical analysis of some ETS equating procedures. In P. Holland (Ed.), Proceedings of the ETS Research Statistics Conference on Test Equating. New York: Academic Press, in press.

Carroll, J. B. The effect of difficulty and chance success on correlations between items or between tests. Psychometrika, 1945, 10, 1-20.

Conrad, L., Trismen, D., & Miller, R. (Eds.) Graduate Record Examinations Technical Manual. Princeton, NJ: Educational Testing Service, 1977.

Cowell, W. ICC preequating in the TOEFL testing program. Paper presented at the meeting of the American Educational Research Association and the National Council on Measurement in Education, San Francisco, April 11, 1979.

Dorans, N., The need for a common metric in item bias studies. U. S. Office of Personnel Management Report TM79-20. Washington, D.C.: U.S. Office of Personnel Management, 1979.

Dressel, P. L. Some remarks on the Kuder-Richardson reliability coefficient. Psychometrika, 1940, 5, 305-310.

Dwyer, P. S. The determination of the factor loadings of a given test from the known factor loadings of other tests. Psychometrika, 1937, 2, 173-178.

Ferguson, G. A. The factorial interpretation of test difficulty. Psychometrika, 1941, 6, 323-329.

Gibson, W. A. Three multivariate models: Factor analysis, latent structure analysis, and latent profile analysis. Psychometrika, 1959, 24, 229-252.

Gibson, W. A. Nonlinear factors in two dimensions. Psychometrika, 1960, 25, 381-392.

Gourlay, N. Difficulty factors arising from the use of tetrachoric correlations in factor analysis. British Journal of Psychology, Statistical Section, 1951, 4, 65-73.

Guilford, J. P. The difficulty of a test and its factor composition. Psychometrika, 1941, 6, 67-77.

Hambleton, R. Latent ability scales: interpretations and uses. In S. Mayo (Ed.), New Directions for Testing and Measurement: Interpreting Test Performance, no. 6. San Francisco: Jossey-Bass, 1980.

Hambleton, R., & Cook, L. Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R., Swaminathan, H., Cook, L., Eignor, D., & Gifford, J. Developments in latent trait theory: Models, technical issues, and applications. Review of Educational Research, 1978, 48, 467-510.

Harman, H. Modern factor analysis (3rd edition). Chicago: University of Chicago Press, 1976.

Jaeger, R. M. Some exploratory indices for selection of a test equating method. Journal of Educational Measurement, 1981, 18, 23-38.

Jennrich, R. I. & Sampson, P. F. Rotation for simple loadings. Psychometrika, 1966, 31, 313-323.

Joreskog, K. G. Structural analysis of covariance and correlation matrices. Psychometrika, 1978, 43, 443-477.

Kaiser, H. F. The varimax criterion for analytical rotation in factor analysis. Psychometrika, 1958, 23, 187-200.

Kingston, N. M. and Dorans, N. J. The effect of the position of an item within a test on item responding behavior: An analysis based on item response theory. Draft report, 1982.

Lord, F. A survey of equating methods based on item characteristic curve theory. Research Bulletin 75-13. Princeton, NJ: Educational Testing Service, 1975a.

Lord, F. Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. Research Bulletin 75-33. Princeton, N. J.: Educational Testing Service, 1975b.

Lord, F. Practical applications of item characteristic curve theory. Journal of Educational Measurement, 1977, 14, 117-138.

Lord, F. Applications of item response theory to practical testing problems. Hillsdale, N.J.: Lawrence Erlbaum Associates, 1980a.

Lord, F., Personal communication, 1980b.

Lord, F., Personal communication, 1981.

Marco, G. Item characteristic curve solutions to three intractable testing problems. Journal of Educational Measurement, 1977, 14, 139-160.

Marco, G., Petersen, N., & Stewart, E. E. A test of the adequacy of curvilinear equating methods. Paper presented at the 1979 Computerized Adaptive Testing Conference, Minneapolis, June 28, 1979.

McDonald, R. P. Nonlinear factor analysis. Psychometric Monographs, 1967, No. 15.

Morris, C. On the foundations of test equating. In P. Holland (Ed.), Proceedings of the ETS Research Statistics Conference on Test Equating. New York: Academic Press, in press.

Mosteller, F., & Tukey, J. Data analysis and regression. Reading, Mass.: Addison-Wesley Publishing Company, 1977.

Petersen, N., Cook, L., & Stocking, M. IRT versus conventional equating methods: A comparative study of scale stability. Paper presented at the meeting of the American Educational Research Association, Los Angeles, April 14, 1981.

Petersen, N., Marco, G., & Stewart, E. E. A test of the adequacy of linear score equating models. In P. Holland (Ed.), Proceedings of the ETS Research Statistics Conference on Test Equating. New York: Academic Press, in press.

Potthoff, R. Some issues in test equating. In P. Holland (Ed.), Proceedings of the ETS Research Statistics Conference on Test Equating. New York: Academic Press, in press.

Powers, D. E., Swinton, S. S., & Carlson, A. B. A factor analytic study of the GRE Aptitude Test. GRE Board Professional Report, GREB No. 75-11P. Princeton, N.J.: Educational Testing Service, 1977.

Powers, D. E., Swinton, S. S., Thayer, D., & Yates, A. A factor analytic study of seven experimental analytical item types. GRE Board Professional Report 77-78. Princeton, NJ: Educational Testing Service, 1978.

Rasch, G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Nielson and Lydicke (for Denmarks Paedagogiske Institut), 1960.

Reckase, M. D. Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 1979, 4, 207-230.

Rock, D., Werts, C., & Grandy, J. Construct validity of the GRE across populations - an empirical confirmatory study. Draft report, 1980.

Stewart, E. E. Equating the Graduate Record Examinations Aptitude Test in the 1980's. Paper submitted to GRE Board Research Committee, April 1981.

Stocking, M. Personal communication, 1980.

Swinton, S. S. Personal communication, 1980.

Swinton, S. S., and Powers, D. E. A factor anlaytic study of the restructured Aptitude Test. GRE Board Professional Report 77-6P. Princeton, N.J.: Educational Testing Service, 1980.

Thurstone, L. L. Multiple common factor analysis. Chicago: University of Chicago Press, 1947.

Tucker, L. R., Koopman, R. F., & Linn, R. L. Evaluation of factor analytic research procedures by means of simulated correlation matrices. Psychometrika, 1969, 34, 421-459.

Warm, T. A primer of item response theory (CG-941278). Oklahoma City: U.S. Coast Guard Institute, December 1978. (NTIS No. AD-A0630).

Wherry, R., and Gaylord, R. Factor pattern of test items and tests as a function of the correlation coefficient: Content, difficulty, and constant error factors. Psychometrika, 1944, 9, 237-244.

Wood, R. L., Wingersky, M., & Lord, F. LOGIST: A computer program for estimating examinee ability and item characteristic curve parameters. ETS Research Memorandum 76-6 (modified 1/78). Princeton, N.J.: Educational Testing Service, 1978.

Wright, B. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Yates, A. An oblique transformation method for primary factor pattern simplification which permits factorial complexity in exploratory analyses. Paper presented at the meeting of the Psychometric Society, Palo Alto, 1974.

Appendix A

Score Conversion Tables for Various Equatings of the Verbal, Quantitative and Analytical Sections of Forms ZGR1, K-ZGR2, K-ZGR3, and 3CGR1

Table A.1

Score Conversion Table for Verbal Scale of Form ZGR1 (2/80)

RAW SCORE	FREQ	IAL	IALH	IALV	LA
80.00	0.0	846.11	846.11	846.11	846.12
79.00	3.00	838.49	838.58	838.57	838.33
78.00	7.00	830.64	831.07	831.00	830.54
77.00	2.00	822.48	823.35	823.13	822.76
76.00	9.00	814.23	815.55	815.10	814.97
75.00	10.00	806.04	807.72	807.02	807.18
74.00	17.00	797.93	799.92	798.96	799.39
73.00	14.00	789.92	792.14	790.93	791.61
72.00	12.00	781.98	784.40	782.94	783.82
71.00	26.00	774.12	776.67	774.99	776.03
70.00	34.00	766.31	768.96	767.08	768.24
69.00	39.00	758.54	761.26	759.19	760.46
68.00	40.00	750.80	753.55	751.32	752.67
67.00	24.00	743.09	745.85	743.47	744.88
66.00	54.00	735.40	738.14	735.62	737.10
65.00	55.00	727.73	730.42	727.77	729.31
64.00	70.00	720.07	722.69	719.92	721.52
63.00	63.00	712.42	714.95	712.08	713.73
62.00	52.00	704.78	707.20	704.23	705.95
61.00	78.00	697.14	699.44	696.38	698.16
60.00	72.00	689.50	691.66	688.54	690.37
59.00	88.00	681.86	683.88	680.69	682.58
58.00	88.00	674.22	676.08	672.86	674.80
57.00	86.00	666.57	668.27	665.03	667.01
56.00	95.00	658.93	660.46	657.21	659.22
55.00	103.00	651.27	652.63	649.41	651.43
54.00	107.00	643.61	644.80	641.62	643.65
53.00	129.00	635.95	636.96	633.85	635.86
52.00	122.00	628.28	629.12	626.11	628.07
51.00	143.00	620.60	621.26	618.39	620.28
50.00	132.00	612.92	613.41	610.69	612.50
49.00	140.00	605.22	605.54	603.02	604.71
48.00	129.00	597.52	597.68	595.38	596.92
47.00	178.00	589.81	589.81	587.75	589.13
46.00	177.00	582.09	581.93	580.15	581.35
45.00	151.00	574.36	574.05	572.58	573.56
44.00	162.00	566.63	566.17	565.01	565.77
43.00	173.00	558.87	558.29	557.47	557.98
42.00	189.00	551.11	550.41	549.94	550.20
41.00	174.00	543.33	542.52	542.42	542.41
40.00	207.00	535.54	534.63	534.90	534.62
39.00	158.00	527.73	526.74	527.39	526.84
38.00	196.00	519.91	518.65	519.88	519.05
37.00	177.00	512.07	510.95	512.37	511.26
36.00	194.00	504.21	503.05	504.85	503.47
35.00	204.00	496.34	495.15	497.33	495.69
34.00	217.00	488.44	487.26	489.79	487.90
33.00	222.00	480.53	479.36	482.25	480.11
32.00	192.00	472.61	471.46	474.70	472.32
31.00	190.00	464.66	463.57	467.14	464.54
30.00	199.00	456.69	455.68	459.57	456.75
29.00	192.00	448.72	447.79	451.98	448.96

Table A.1 continued

Score Conversion Table for Verbal Scale of Form ZGR1 (2/80)

28.00	189.00	440.72	439.90	444.39	441.17
27.00	173.00	432.71	432.02	436.78	433.39
26.00	187.00	424.69	424.14	429.17	425.60
25.00	170.00	416.65	416.27	421.54	417.81
24.00	153.00	408.60	408.40	413.90	410.02
23.00	148.00	400.54	400.54	406.25	402.24
22.00	157.00	392.46	392.69	398.60	394.45
21.00	136.00	384.38	384.84	390.93	386.66
20.00	130.00	376.29	377.00	383.24	378.87
19.00	121.00	368.19	369.16	375.55	371.09
18.00	113.00	360.09	361.33	367.84	363.30
17.00	93.00	351.98	353.51	360.12	355.51
16.00	107.00	343.87	345.70	352.38	347.73
15.00	96.00	335.77	337.89	344.63	339.94
14.00	109.00	327.66	330.10	336.86	332.15
13.00	82.00	319.57	322.31	329.06	324.36
12.00	59.00	311.49	314.54	321.25	316.58
11.00	67.00	303.42	306.78	313.41	308.79
10.00	66.00	295.38	299.04	305.54	301.00
9.00	58.00	287.36	291.32	297.64	293.21
8.00	42.00	279.37	283.61	289.70	285.43
7.00	46.00	271.41	275.91	281.73	277.64
6.00	51.00	263.49	268.23	273.73	269.85
5.00	38.00	255.63	260.56	265.69	262.06
4.00	50.00	247.81	252.89	257.61	254.28
3.00	38.00	240.04	245.23	249.51	246.49
2.00	21.00	232.33	237.55	241.39	238.70
1.00	37.00	224.67	229.86	233.25	230.91
0.0	28.00	217.06	222.13	225.08	223.13
-1.00	15.00	209.45	214.35	216.86	215.34
-2.00	19.00	201.42	206.42	208.45	207.55
-3.00	7.00	193.76	198.35	200.13	199.76
-4.00	9.00	186.11	190.60	192.36	191.98
-5.00	3.00	178.45	182.86	184.60	184.19
-6.00	4.00	170.80	175.11	176.84	176.40
-7.00	1.00	163.14	167.37	169.07	168.61
-8.00	1.00	155.49	159.62	161.31	160.83
-9.00	1.00	147.83	151.88	153.55	153.04
-10.00	0.0	140.18	144.13	145.79	145.25
-11.00	0.0	132.52	136.39	138.02	137.47
-12.00	1.00	174.86	128.64	130.26	129.68

Table A.2

Score Conversion Table for Verbal Scale of
Form K-ZGR2

RAW SCORE	FREQ	IAL	IAW	IALH	IALV	LA
80.00	0.0	846.11	846.11	846.11	846.11	846.62
79.00	1.00	839.75	840.24	839.09	839.07	838.69
78.00	5.00	830.95	831.81	829.76	829.67	830.76
77.00	5.00	821.52	823.13	819.87	819.55	822.83
76.00	7.00	812.17	814.72	810.16	809.54	814.91
75.00	14.00	803.06	806.62	800.80	799.86	806.98
74.00	25.00	794.21	798.75	791.77	790.53	799.05
73.00	18.00	785.60	791.04	783.04	781.54	791.12
72.00	19.00	777.16	783.45	774.57	772.83	783.19
71.00	26.00	768.89	775.92	766.30	764.35	775.27
70.00	22.00	760.74	768.43	758.19	756.06	767.34
69.00	33.00	752.69	760.96	750.23	747.93	759.41
68.00	39.00	744.72	753.47	742.39	739.94	751.48
67.00	38.00	736.82	745.96	734.63	732.05	743.55
66.00	41.00	728.97	738.41	726.94	724.24	735.62
65.00	51.00	721.16	730.83	719.31	716.50	727.70
64.00	53.00	713.37	723.20	711.72	708.81	719.77
63.00	64.00	705.60	715.53	704.15	701.15	711.84
62.00	61.00	697.83	707.80	696.60	693.52	703.91
61.00	62.00	690.06	700.02	689.05	685.91	695.98
60.00	79.00	682.28	692.18	681.49	678.30	688.06
59.00	106.00	674.49	684.28	673.92	670.69	680.13
58.00	102.00	666.68	676.33	666.34	663.09	672.20
57.00	93.00	658.85	668.33	658.72	655.48	664.27
56.00	101.00	651.00	660.27	651.08	647.86	656.34
55.00	113.00	643.13	652.16	643.40	640.23	648.42
54.00	135.00	635.23	644.01	635.68	632.59	640.49
53.00	132.00	627.31	635.81	627.93	624.94	632.56
52.00	140.00	619.37	627.55	620.14	617.28	624.63
51.00	128.00	611.41	619.30	612.31	609.62	616.70
50.00	146.00	603.43	611.03	604.44	601.94	608.78
49.00	140.00	595.43	602.73	596.54	594.27	600.85
48.00	150.00	587.43	594.42	588.60	586.59	592.92
47.00	171.00	579.41	586.11	580.64	578.91	584.99
46.00	158.00	571.39	577.79	572.66	571.24	577.06
45.00	193.00	563.36	569.49	564.66	563.57	569.13
44.00	190.00	555.33	561.19	556.65	555.90	561.21
43.00	172.00	547.30	552.92	548.63	548.24	553.28
42.00	187.00	539.27	544.67	540.60	540.59	545.35
41.00	201.00	531.24	536.45	532.57	532.95	537.42
40.00	179.00	523.21	528.26	524.56	525.32	529.49
39.00	187.00	515.18	520.10	516.55	517.70	521.57
38.00	212.00	507.16	511.98	508.56	510.10	513.64
37.00	220.00	499.14	503.90	500.59	502.51	505.71
36.00	202.00	491.13	495.86	492.65	494.94	497.78
35.00	187.00	483.12	487.83	484.73	487.38	489.85
34.00	215.00	475.12	479.85	476.84	479.85	481.93
33.00	209.00	467.13	471.91	468.98	472.33	474.00
32.00	204.00	459.15	463.99	461.15	464.82	466.07
31.00	181.00	451.18	456.10	453.35	457.34	458.14
30.00	175.00	443.22	448.24	445.58	449.86	450.21
29.00	221.00	435.27	440.41	437.84	442.40	442.28

Table A.2 continued

Score Conversion Table for Verbal Scale of
Form K-ZGR2

28.00	171.00	427.34	432.60	430.12	434.95	434.36
27.00	190.00	419.42	424.81	422.43	427.51	426.43
26.00	166.00	411.53	417.06	414.76	420.07	418.50
25.00	191.00	403.66	409.35	407.11	412.64	410.57
24.00	169.00	395.83	401.67	399.48	405.22	402.64
23.00	134.00	388.02	394.04	391.88	397.80	394.72
22.00	128.00	380.26	386.46	384.29	390.39	386.79
21.00	128.00	372.54	378.92	376.73	382.98	378.86
20.00	126.00	364.86	371.45	369.19	375.58	370.93
19.00	106.00	357.24	364.04	361.68	368.19	363.00
18.00	116.00	349.68	356.69	354.21	360.81	355.08
17.00	119.00	342.17	349.40	346.77	353.45	347.15
16.00	98.00	334.73	342.17	339.37	346.10	339.22
15.00	80.00	327.35	335.00	332.03	338.79	331.29
14.00	71.00	320.03	327.87	324.74	331.49	323.36
13.00	69.00	312.78	320.77	317.50	324.23	315.44
12.00	76.00	305.58	313.71	310.33	316.99	307.51
11.00	46.00	298.44	306.65	303.22	309.79	299.58
10.00	49.00	291.34	299.61	296.17	302.60	291.65
9.00	53.00	284.29	292.56	289.18	295.45	283.72
8.00	65.00	277.27	285.50	282.25	288.31	275.79
7.00	36.00	270.28	278.43	275.38	281.17	267.87
6.00	39.00	263.29	271.32	268.54	274.05	259.94
5.00	43.00	256.30	264.19	261.73	266.91	252.01
4.00	26.00	249.30	257.01	254.93	259.76	244.08
3.00	30.00	242.27	249.79	248.12	252.57	236.15
2.00	32.00	235.21	242.52	241.29	245.35	228.23
1.00	16.00	228.11	235.23	234.42	238.08	220.30
0.0	19.00	220.98	227.89	227.50	230.76	212.37
-1.00	13.00	213.83	220.51	220.54	223.41	204.44
-2.00	19.00	206.38	212.79	213.57	216.04	196.51
-3.00	6.00	198.46	205.60	206.60	208.64	188.59
-4.00	10.00	190.71	197.69	198.79	200.57	180.66
-5.00	4.00	182.96	189.78	190.81	192.57	172.73
-6.00	3.00	175.21	181.87	182.83	184.57	164.80
-7.00	3.00	167.46	173.96	174.85	176.57	156.87
-8.00	1.00	159.71	166.05	166.87	168.57	148.95
-9.00	0.0	151.96	158.14	158.89	160.57	141.02
-10.00	0.0	144.21	150.23	150.91	152.58	133.09

Table A.3 continued

Score Conversion Table for Verbal Scale of
Form K-ZGR3

28.00	447.00	422.18	419.70	421.47	420.44	425.58	422.77	425.97
27.00	427.00	413.71	411.42	413.02	412.17	417.56	414.57	417.49
26.00	426.00	405.27	403.19	404.59	403.95	409.57	406.04	409.00
25.00	410.00	396.84	395.02	396.18	395.77	401.60	396.97	400.52
24.00	390.00	388.45	386.92	387.80	387.63	393.65	387.93	392.04
23.00	395.00	380.09	378.89	379.46	379.53	385.72	378.93	383.56
22.00	280.00	371.79	370.95	371.17	371.48	377.82	370.12	375.07
21.00	280.00	363.55	363.10	362.95	363.48	369.96	362.04	366.59
20.00	265.00	355.38	355.33	354.80	355.54	362.13	353.84	358.11
19.00	202.00	347.30	347.66	346.74	347.67	354.34	346.13	349.62
18.00	214.00	339.32	340.09	338.77	339.89	346.61	338.80	341.14
17.00	195.00	331.44	332.60	330.91	332.19	338.95	331.54	332.66
16.00	168.00	323.67	325.22	323.16	324.59	331.35	324.45	324.17
15.00	163.00	316.02	317.92	315.54	317.11	323.83	316.88	315.69
14.00	153.00	305.50	310.72	308.04	309.74	316.40	308.59	307.21
13.00	143.00	301.10	303.60	300.66	302.50	309.06	300.62	298.72
12.00	119.00	293.83	296.58	293.42	295.39	301.81	293.34	290.24
11.00	115.00	286.68	289.63	286.29	288.41	294.66	285.68	281.76
10.00	113.00	279.66	282.79	279.29	281.56	287.59	276.77	273.27
9.00	97.00	272.74	276.03	272.40	274.83	280.61	269.07	264.79
8.00	88.00	265.93	269.36	265.61	268.22	273.71	262.43	256.31
7.00	58.00	259.22	262.78	258.94	261.71	266.89	255.99	247.82
6.00	68.00	252.62	256.28	252.36	255.29	260.14	249.63	239.34
5.00	54.00	246.11	249.86	245.88	248.96	253.46	242.53	230.86
4.00	48.00	239.70	243.53	239.51	242.70	246.84	235.84	222.37
3.00	53.00	233.41	237.30	233.25	236.52	240.30	228.60	213.89
2.00	29.00	227.26	231.19	227.13	230.42	233.84	221.85	205.41
1.00	30.00	221.29	225.27	221.20	224.41	227.50	214.87	196.92
0.0	26.00	215.58	219.66	215.52	218.56	221.32	205.36	188.44
-1.00	11.00	210.25	214.70	210.22	212.94	215.38	194.79	179.96
-2.00	7.00	205.14	210.09	205.14	207.72	209.83	186.83	171.48
-3.00	4.00	197.40	202.24	197.40	202.46	204.25	180.88	162.99
-4.00	2.00	189.66	194.40	189.66	194.62	196.40	177.00	154.51
-5.00	5.00	181.92	186.56	181.92	186.78	188.54	170.13	146.03
-6.00	0.0	174.18	178.72	174.18	178.95	180.68	161.31	137.54
-7.00	2.00	166.44	170.88	166.44	171.11	172.82	140.01	129.06
-8.00	0.0	158.70	163.04	158.70	163.27	164.96	140.01	120.58
-9.00	0.0	150.96	155.20	150.96	155.43	157.11	140.01	112.09
-10.00	0.0	143.22	147.36	143.22	147.59	149.25	140.01	103.61

Table A.3

Score Conversion Table for Verbal Scale of
Form K-ZGR3

RAW SCORE	FREQ	IAL	IAW	IAL2	IALH	IALV	EE	LE
80.00	1.00	846.11	846.11	846.11	846.11	846.11	845.61	867.10
79.00	1.00	838.93	840.06	838.89	839.65	839.63	841.47	858.61
78.00	13.00	831.44	833.06	831.36	833.13	833.09	832.47	850.13
77.00	1.00	824.29	826.49	824.17	826.70	826.55	826.28	841.65
76.00	15.00	817.30	820.12	817.13	820.16	819.85	819.52	833.16
75.00	28.00	810.33	813.78	810.12	813.42	812.91	808.96	824.68
74.00	31.00	803.30	807.33	803.05	806.47	805.73	801.68	816.20
73.00	46.00	796.16	800.70	795.88	799.29	798.31	794.61	807.72
72.00	21.00	788.90	793.86	788.58	791.90	790.67	788.87	799.23
71.00	53.00	781.50	786.80	781.14	784.32	782.86	782.56	790.75
70.00	60.00	773.97	779.50	773.57	776.57	774.88	774.25	782.27
69.00	78.00	766.31	771.99	765.87	768.68	766.79	766.57	773.78
68.00	102.00	758.55	764.29	758.07	760.68	758.61	758.13	765.30
67.00	84.00	750.69	756.42	750.18	752.62	750.37	750.66	756.82
66.00	99.00	742.76	748.40	742.22	744.49	742.08	744.06	748.33
65.00	111.00	734.76	740.25	734.19	736.33	733.78	737.08	739.85
64.00	129.00	726.71	732.01	726.11	728.14	725.45	729.90	731.37
63.00	141.00	718.60	723.67	717.98	719.92	717.12	722.33	722.88
62.00	157.00	710.46	715.26	709.81	711.69	708.78	714.57	714.40
61.00	181.00	702.27	706.79	701.60	703.44	700.44	705.83	705.92
60.00	175.00	694.04	698.27	693.35	695.17	692.08	697.46	697.43
59.00	199.00	685.77	689.69	685.07	686.88	683.72	689.95	688.95
58.00	223.00	677.46	681.07	676.74	678.54	675.33	682.38	680.47
57.00	193.00	669.12	672.40	668.37	670.18	666.94	675.40	671.98
56.00	249.00	660.73	663.69	659.97	661.78	658.53	668.05	663.50
55.00	259.00	652.30	654.94	651.53	653.34	650.11	659.82	655.02
54.00	329.00	643.84	646.15	643.06	644.85	641.67	651.18	646.53
53.00	314.00	635.35	637.33	634.55	636.32	633.22	641.90	638.05
52.00	311.00	626.83	628.47	626.02	627.74	624.75	633.12	629.57
51.00	336.00	618.28	619.61	617.47	619.12	616.28	624.40	621.08
50.00	395.00	609.72	610.72	608.90	610.45	607.81	615.19	612.60
49.00	427.00	601.15	601.83	600.33	601.76	599.34	605.75	604.12
48.00	399.00	592.57	592.92	591.74	593.03	590.87	596.97	595.64
47.00	418.00	583.99	584.02	583.16	584.28	582.42	588.73	587.15
46.00	470.00	575.41	575.13	574.58	575.51	573.97	580.06	578.67
45.00	423.00	566.84	566.25	566.01	566.72	565.54	571.82	570.19
44.00	494.00	558.28	557.38	557.45	557.93	557.13	563.49	561.70
43.00	528.00	549.72	548.54	548.90	549.14	548.73	554.31	553.22
42.00	445.00	541.18	539.72	540.36	540.36	540.36	545.69	544.74
41.00	504.00	532.64	530.93	531.83	531.58	532.00	537.64	536.25
40.00	567.00	524.12	522.17	523.31	522.82	523.67	529.02	527.77
39.00	551.00	515.60	513.44	514.80	514.09	515.36	520.41	519.29
38.00	540.00	507.09	504.74	506.29	505.38	507.07	511.86	510.80
37.00	525.00	498.59	496.08	497.80	496.71	498.80	503.28	502.32
36.00	573.00	490.09	487.45	489.30	488.07	490.57	494.76	493.84
35.00	596.00	481.59	478.86	480.82	479.46	482.35	485.47	485.35
34.00	541.00	473.10	470.30	472.33	470.90	474.16	476.28	476.87
33.00	582.00	464.60	461.77	463.85	462.38	466.00	467.14	468.39
32.00	518.00	456.11	453.27	455.36	453.91	457.87	457.90	459.90
31.00	528.00	447.62	444.82	446.89	445.47	449.76	448.76	451.42
30.00	476.00	439.14	436.40	438.41	437.09	441.68	439.63	442.94
29.00	447.00	430.65	428.02	429.94	428.74	433.62	431.00	434.45

Table A.4

Score Conversion Table for Verbal Scale of Form 3CGR1

RAW SCORES	FREQ	IEP	IES	IESH	IESV	EE	LE
75.00	1.00	846.11	846.11	846.11	846.11	836.46	829.71
74.00	3.00	838.20	838.90	838.01	837.99	829.66	821.70
73.00	4.00	829.83	829.68	828.25	828.13	822.42	813.69
72.00	2.00	821.72	820.48	818.87	818.53	818.69	805.68
71.00	7.00	813.80	811.53	809.95	809.32	813.94	797.67
70.00	11.00	805.94	802.76	801.30	800.38	807.83	789.65
69.00	20.00	798.05	794.07	792.82	791.62	798.34	781.64
68.00	28.00	790.06	785.43	784.42	782.96	785.78	773.63
67.00	15.00	781.95	776.78	776.07	774.37	775.83	765.62
66.00	21.00	773.69	768.13	767.73	765.82	769.98	757.61
65.00	48.00	765.29	759.47	759.41	757.30	761.44	749.59
64.00	49.00	756.76	750.82	751.10	748.82	752.67	741.58
63.00	56.00	748.13	742.19	742.80	740.36	741.88	733.57
62.00	37.00	739.42	733.59	734.52	731.94	734.23	725.56
61.00	85.00	730.63	725.02	726.25	723.54	726.67	717.55
60.00	79.00	721.81	716.48	718.01	715.18	717.85	709.53
59.00	74.00	712.96	707.98	709.77	706.84	709.86	701.52
58.00	100.00	704.10	699.51	701.55	698.52	701.10	693.51
57.00	70.00	695.24	691.08	693.34	690.23	693.87	685.50
56.00	127.00	686.39	682.67	685.13	681.96	686.32	677.49
55.00	114.00	677.55	674.28	676.92	673.70	677.26	669.47
54.00	131.00	668.73	665.90	668.70	665.46	667.91	661.46
53.00	148.00	659.93	657.54	660.47	657.22	658.89	653.45
52.00	147.00	651.16	649.19	652.22	648.99	649.56	645.44
51.00	170.00	642.42	640.85	643.95	640.77	640.53	637.43
50.00	186.00	633.70	632.51	635.65	632.56	631.67	629.41
49.00	189.00	625.01	624.17	627.32	624.34	622.24	621.40
48.00	202.00	616.35	615.83	618.96	616.13	613.54	613.39
47.00	215.00	607.72	607.50	610.57	607.93	604.95	605.38
46.00	248.00	599.13	599.17	602.15	599.72	595.39	597.37
45.00	209.00	590.57	590.83	593.70	591.52	586.80	589.35
44.00	222.00	582.04	582.51	585.22	583.33	579.41	581.34
43.00	273.00	573.55	574.18	576.73	575.14	571.15	573.33
42.00	254.00	565.11	565.87	568.21	566.97	562.81	565.32
41.00	284.00	556.70	557.56	559.68	558.80	554.59	557.31
40.00	325.00	548.34	549.26	551.14	550.64	545.57	549.29
39.00	304.00	540.02	540.98	542.61	542.51	536.83	541.28
38.00	325.00	531.75	532.71	534.09	534.38	528.89	533.27
37.00	294.00	523.53	524.46	525.57	526.28	521.20	525.26
36.00	326.00	515.36	516.24	517.08	518.20	513.28	517.25
35.00	337.00	507.23	508.04	508.62	510.15	505.05	509.23
34.00	330.00	499.14	499.87	500.19	502.13	496.97	501.22
33.00	355.00	491.10	491.72	491.80	494.13	489.00	493.21
32.00	339.00	483.10	483.60	483.45	496.16	481.21	485.20
31.00	345.00	475.13	475.51	475.14	478.22	473.47	477.19
30.00	357.00	467.20	467.44	466.87	470.30	465.50	469.17
29.00	359.00	459.29	459.39	458.64	462.41	457.62	461.16
28.00	343.00	451.41	451.37	450.44	454.54	449.85	453.15
27.00	352.00	443.55	443.36	442.29	446.69	442.37	445.14
26.00	330.00	435.70	435.37	434.16	438.86	434.91	437.13
25.00	364.00	427.85	427.39	426.07	431.03	426.84	429.11
24.00	340.00	420.01	419.42	417.99	423.21	418.98	421.10

Table A.4 continued

Score Conversion Table for Verbal Scale of
Form 3CGR1

23.00	303.00	412.17	411.46	409.94	415.40	411.91	413.09
22.00	332.00	404.32	403.51	401.91	407.58	404.27	405.08
21.00	287.00	396.46	395.56	393.89	399.77	396.35	397.07
20.00	311.00	388.59	387.62	385.88	391.94	388.60	389.05
19.00	289.00	380.71	379.68	377.89	384.12	380.51	381.04
18.00	263.00	372.82	371.75	369.90	376.28	372.87	373.03
17.00	242.00	364.91	363.83	361.93	368.43	365.93	365.02
16.00	242.00	357.00	355.92	353.97	360.57	359.21	357.01
15.00	226.00	349.07	348.01	346.07	352.70	352.25	348.99
14.00	222.00	341.13	340.10	338.08	344.82	344.76	340.98
13.00	194.00	333.19	332.19	330.15	336.91	337.41	332.97
12.00	198.00	325.22	324.27	322.24	328.99	330.52	324.96
11.00	172.00	317.24	316.33	314.33	321.04	323.51	316.95
10.00	184.00	309.24	308.37	306.44	313.06	315.66	308.93
9.00	159.00	301.20	300.38	298.55	305.04	308.07	300.92
8.00	176.00	293.12	292.35	290.67	296.97	300.57	292.91
7.00	154.00	285.00	284.28	282.79	288.86	292.12	284.90
6.00	152.00	276.84	276.17	274.91	280.69	283.28	276.89
5.00	153.00	268.64	268.02	267.04	272.48	272.78	268.87
4.00	115.00	260.43	259.84	259.16	264.22	263.04	260.86
3.00	107.00	252.21	251.65	251.30	255.93	255.38	252.85
2.00	75.00	244.02	243.48	243.44	247.62	248.85	244.84
1.00	99.00	235.88	235.34	235.62	239.35	241.06	236.83
0.0	74.00	228.08	227.26	227.86	231.14	229.66	228.81
-1.00	47.00	220.39	219.22	220.18	223.03	219.98	220.80
-2.00	36.00	212.95	211.05	212.57	214.97	211.56	212.79
-3.00	14.00	205.69	202.11	204.31	206.16	203.03	204.78
-4.00	14.00	197.49	194.01	196.01	197.79	195.35	196.77
-5.00	7.00	189.29	185.90	187.84	189.59	186.32	188.75
-6.00	6.00	181.09	177.80	179.66	181.40	176.70	180.74
-7.00	3.00	172.89	169.70	171.49	173.21	164.17	172.73
-8.00	2.00	164.69	161.59	163.32	165.02	152.72	164.72

Table A.5

Score Conversion Table for Quantitative Scale of Form ZGR1 (2/80)

RAW SCORES	FREQ	IAL	LA
55.00	3.00	883.14	883.15
54.00	6.00	870.44	870.49
53.00	5.00	857.91	857.82
52.00	19.00	845.58	845.16
51.00	27.00	833.30	832.49
50.00	32.00	820.96	819.83
49.00	47.00	808.50	807.16
48.00	34.00	795.94	794.50
47.00	72.00	783.29	781.83
46.00	70.00	770.58	769.16
45.00	109.00	757.86	756.50
44.00	73.00	745.13	743.83
43.00	105.00	732.42	731.17
42.00	121.00	719.73	718.50
41.00	147.00	707.07	705.84
40.00	137.00	694.44	693.17
39.00	170.00	681.85	680.51
38.00	176.00	669.29	667.84
37.00	202.00	656.76	655.18
36.00	207.00	644.27	642.51
35.00	214.00	631.79	629.84
34.00	251.00	619.32	617.18
33.00	240.00	606.86	604.51
32.00	281.00	594.40	591.85
31.00	273.00	581.93	579.18
30.00	319.00	569.44	566.52
29.00	312.00	556.94	553.85
28.00	306.00	544.40	541.19
27.00	331.00	531.83	528.52
26.00	341.00	519.23	515.85
25.00	317.00	506.57	503.19
24.00	304.00	493.86	490.52
23.00	311.00	481.09	477.86
22.00	285.00	468.24	465.19
21.00	270.00	455.33	452.53
20.00	277.00	442.36	439.86
19.00	273.00	429.34	427.20
18.00	236.00	416.28	414.53
17.00	207.00	403.20	401.87
16.00	189.00	390.14	389.20
15.00	150.00	377.10	376.53
14.00	139.00	364.11	363.87
13.00	109.00	351.18	351.20
12.00	120.00	338.31	338.54
11.00	106.00	325.51	325.87
10.00	102.00	312.75	313.21
9.00	68.00	300.04	300.54
8.00	48.00	287.37	287.88
7.00	48.00	274.74	275.21
6.00	53.00	262.14	262.54
5.00	55.00	249.57	249.88
4.00	33.00	237.06	237.21

Table A.5 continued

Score Conversion Table for Quantitative Scale of Form ZGR1 (2/80)

3.00	32.00	224.60	224.55
2.00	37.00	212.20	211.88
1.00	24.00	199.95	199.22
0.0	33.00	187.53	186.55
-1.00	8.00	175.17	173.89
-2.00	8.00	162.71	161.22
-3.00	4.00	150.01	148.56
-4.00	2.00	137.15	135.89
-5.00	3.00	124.44	123.22
-6.00	0.0	111.74	110.56
-7.00	1.00	99.03	97.89
-8.00	0.0	86.32	85.23
-9.00	0.0	73.62	72.56
-10.00	0.0	60.91	59.90

Table A.6

Score Conversion Table for Quantitative Scale of Form K-ZGR2

RAW SCORES	FREQ	IAL	LA
55.00	5.00	883.14	867.30
54.00	16.00	861.69	854.87
53.00	6.00	845.47	842.45
52.00	27.00	830.89	830.03
51.00	36.00	816.77	817.60
50.00	60.00	802.76	805.18
49.00	60.00	788.83	792.75
48.00	48.00	775.05	780.33
47.00	88.00	761.50	767.91
46.00	97.00	748.23	755.48
45.00	115.00	735.24	743.06
44.00	74.00	722.52	730.64
43.00	113.00	710.04	718.21
42.00	136.00	697.75	705.79
41.00	177.00	685.62	693.37
40.00	142.00	673.60	680.94
39.00	169.00	661.68	668.52
38.00	208.00	649.82	656.09
37.00	220.00	638.04	643.67
36.00	215.00	626.32	631.25
35.00	216.00	614.70	618.82
34.00	246.00	603.19	606.40
33.00	268.00	591.81	593.98
32.00	268.00	580.58	581.55
31.00	271.00	569.49	569.13
30.00	290.00	558.55	556.70
29.00	297.00	547.74	544.28
28.00	327.00	537.03	531.86
27.00	277.00	526.39	519.43
26.00	298.00	515.77	507.01
25.00	337.00	505.11	494.59
24.00	308.00	494.36	482.16
23.00	285.00	483.45	469.74
22.00	258.00	472.33	457.31
21.00	266.00	460.94	444.89
20.00	260.00	449.23	432.47
19.00	233.00	437.16	420.04
18.00	216.00	424.70	407.62
17.00	224.00	411.85	395.20
16.00	197.00	398.62	382.77
15.00	146.00	385.04	370.35
14.00	148.00	371.16	357.93
13.00	127.00	357.03	345.50
12.00	96.00	342.70	333.08
11.00	95.00	328.22	320.65
10.00	90.00	313.65	308.23
9.00	66.00	299.03	295.81
8.00	57.00	284.44	283.38
7.00	66.00	269.97	270.96
6.00	38.00	255.75	258.54
5.00	44.00	241.93	246.11
4.00	30.00	228.65	233.69

Table A.6 continued

Score Conversion Table for Quantitative Scale of Form K-ZGR2

3.00	35.00	216.06	221.26
2.00	29.00	204.27	208.84
1.00	17.00	193.34	196.42
0.0	11.00	183.24	183.99
-1.00	5.00	173.93	171.57
-2.00	2.00	165.29	159.15
-3.00	2.00	157.09	146.72
-4.00	3.00	148.06	134.30
-5.00	0.0	134.24	121.88
-6.00	1.00	121.08	109.45
-7.00	0.0	107.93	97.03
-8.00	0.0	94.77	84.60
-9.00	0.0	81.61	72.18
-10.00	0.0	68.45	59.76

Table A.7

Score Conversion Table for Quantitative Scale of Form K-ZGR3

RAW SCORES	FREQ	IAL	IAL2	LE
55.00	4.00	883.14	883.14	842.40
54.00	6.00	848.18	849.67	830.81
53.00	3.00	828.04	830.17	819.21
52.00	10.00	811.64	814.26	807.61
51.00	14.00	796.78	799.75	796.02
50.00	25.00	782.73	785.94	784.42
49.00	21.00	769.17	772.54	772.82
48.00	29.00	755.99	759.42	761.23
47.00	38.00	743.14	746.57	749.63
46.00	43.00	730.59	733.98	738.03
45.00	42.00	718.33	721.64	726.44
44.00	52.00	706.34	709.55	714.84
43.00	46.00	694.61	697.69	703.24
42.00	66.00	683.10	686.05	691.65
41.00	64.00	671.80	674.60	680.05
40.00	64.00	660.68	663.33	668.45
39.00	87.00	649.72	652.20	656.86
38.00	103.00	638.91	641.22	645.26
37.00	98.00	628.22	630.35	633.66
36.00	104.00	617.66	619.61	622.07
35.00	86.00	607.21	608.97	610.47
34.00	112.00	596.87	598.44	598.87
33.00	124.00	586.63	588.01	587.28
32.00	120.00	576.49	577.68	575.68
31.00	110.00	566.42	567.42	564.08
30.00	125.00	556.42	557.22	552.49
29.00	154.00	546.44	547.07	540.89
28.00	150.00	536.47	536.91	529.29
27.00	146.00	526.46	526.72	517.70
26.00	160.00	516.38	516.46	506.10
25.00	161.00	506.18	506.09	494.50
24.00	152.00	495.83	495.55	482.91
23.00	150.00	485.28	484.82	471.31
22.00	145.00	474.50	473.85	459.71
21.00	166.00	463.48	462.63	448.12
20.00	145.00	452.20	451.15	436.52
19.00	138.00	440.68	439.42	424.92
18.00	146.00	428.94	427.48	413.33
17.00	142.00	417.04	415.38	401.73
16.00	140.00	405.01	403.16	390.14
15.00	120.00	392.93	390.90	378.54
14.00	121.00	380.84	378.64	366.94
13.00	92.00	368.78	366.42	355.35
12.00	90.00	356.76	354.26	343.75
11.00	88.00	344.79	342.16	332.15
10.00	63.00	332.85	330.11	320.56
9.00	73.00	320.91	318.08	308.96
8.00	63.00	308.92	306.01	297.36
7.00	49.00	296.85	293.88	285.77
6.00	44.00	284.65	281.64	274.17
5.00	30.00	272.27	269.26	262.57
4.00	21.00	259.72	256.73	250.98

Table A.7 continued

Score Conversion Table for Quantitative Scale of Form K-ZGR3

3.00	31.00	246.97	244.07	239.38
2.00	15.00	234.08	231.31	227.78
1.00	17.00	221.07	218.51	216.19
0.0	9.00	208.00	205.71	204.59
-1.00	6.00	194.88	192.93	192.99
-2.00	3.00	181.58	180.05	181.40
-3.00	2.00	167.76	166.72	169.80
-4.00	0.0	152.49	152.11	158.20
-5.00	0.0	137.05	137.05	146.61
-6.00	0.0	123.97	123.97	135.01
-7.00	0.0	110.85	110.88	123.41
-8.00	0.0	97.79	97.79	111.82
-9.00	0.0	84.70	84.70	100.22
-10.00	0.0	71.61	71.61	88.62

Table A.8

Score Conversion Table for Quantitative Scale of Form 3CGR1

RAW SCORE	FREQ	IEP	IES	EE	LE
55.00	9.00	883.14	883.14	877.82	837.06
54.00	21.00	859.86	861.32	849.13	825.74
53.00	20.00	839.70	841.12	836.86	814.43
52.00	38.00	820.64	822.57	824.90	803.11
51.00	62.00	801.85	804.59	807.79	791.79
50.00	103.00	783.51	787.07	785.65	780.47
49.00	93.00	765.92	770.18	770.90	769.16
48.00	79.00	749.29	754.08	759.79	757.84
47.00	122.00	733.66	738.80	745.87	746.52
46.00	137.00	718.95	724.33	729.68	735.20
45.00	161.00	705.05	710.58	715.78	723.88
44.00	146.00	691.84	697.45	704.01	712.57
43.00	183.00	679.21	684.85	691.87	701.25
42.00	203.00	667.05	672.69	678.17	689.93
41.00	224.00	655.30	660.90	665.49	678.61
40.00	238.00	643.89	649.43	653.47	667.30
39.00	233.00	632.79	638.25	642.51	655.98
38.00	238.00	621.96	627.33	632.43	644.66
37.00	252.00	611.38	616.65	623.05	633.34
36.00	305.00	601.04	606.21	613.28	622.03
35.00	278.00	590.93	595.99	603.76	610.71
34.00	321.00	581.03	585.97	594.49	599.39
33.00	322.00	571.32	576.16	584.89	588.07
32.00	354.00	561.78	566.52	575.16	576.75
31.00	387.00	552.39	557.04	564.95	565.44
30.00	419.00	543.12	547.69	554.57	554.12
29.00	427.00	533.94	538.43	544.15	542.80
28.00	424.00	524.80	529.23	533.96	531.48
27.00	445.00	515.68	520.05	523.70	520.17
26.00	449.00	506.54	510.85	513.28	508.85
25.00	487.00	497.34	501.59	502.68	497.53
24.00	506.00	488.03	492.22	492.13	486.21
23.00	475.00	478.58	482.70	481.69	474.90
22.00	441.00	468.95	473.00	471.70	463.58
21.00	456.00	459.11	463.06	461.75	452.26
20.00	463.00	449.02	452.87	451.47	440.94
19.00	460.00	438.66	442.39	440.56	429.62
18.00	466.00	428.01	431.60	429.21	418.31
17.00	411.00	417.04	420.51	417.92	406.99
16.00	450.00	405.76	409.11	405.68	395.67
15.00	411.00	394.15	397.44	392.22	384.35
14.00	357.00	382.24	385.52	378.75	373.04
13.00	310.00	370.03	373.39	365.56	361.72
12.00	304.00	357.54	361.11	351.34	350.40
11.00	263.00	344.79	348.70	335.89	339.08
10.00	256.00	331.79	336.20	321.65	327.77
9.00	229.00	318.55	323.63	306.89	316.45
8.00	196.00	305.08	310.99	291.77	305.13
7.00	159.00	291.39	298.28	276.45	293.81
6.00	131.00	277.49	285.48	262.43	282.49
5.00	106.00	263.45	272.59	249.26	271.18
4.00	84.00	249.34	259.61	235.41	259.86

Table A.8 continued

Score Conversion Table for Quantitative Scale of Form 3CGR1

3.00	77.00	235.26	246.56	221.81	248.54
2.00	50.00	221.35	233.51	210.46	237.22
1.00	63.00	207.73	220.48	198.16	225.91
0.0	37.00	194.48	207.54	180.61	214.59
-1.00	24.00	181.67	194.65	167.13	203.27
-2.00	26.00	169.28	181.64	153.21	191.95
-3.00	12.00	157.26	167.94	129.36	180.64
-4.00	3.00	146.65	151.89	115.19	169.32
-5.00	4.00	134.82	138.43	106.35	158.00
-6.00	2.00	121.86	125.22	91.05	146.68
-7.00	0.0	108.90	112.02	91.05	135.36
-8.00	0.0	95.94	98.81	91.05	124.09
-9.00	0.0	82.98	85.61	91.05	112.73
-10.00	0.0	70.03	72.40	91.05	101.41

Table A.9

Score Conversion Table for Analytical Scale of Form 3CGR1

RAW SCOPE	FREQ	IEP	IES	EE	LE
66.00	0.0	805.71	805.71	797.55	813.48
65.00	1.00	790.62	793.48	797.55	804.81
64.00	7.00	777.32	782.63	773.61	796.14
63.00	1.00	764.85	772.58	769.43	787.48
62.00	11.00	752.96	762.96	765.35	778.81
61.00	17.00	741.55	753.65	758.86	770.14
60.00	30.00	730.56	744.55	751.07	761.48
59.00	68.00	719.96	735.62	735.78	752.81
58.00	41.00	709.66	726.84	726.78	744.14
57.00	83.00	699.64	718.17	720.46	735.48
56.00	76.00	689.83	709.61	714.39	726.81
55.00	109.00	680.20	701.13	708.52	718.14
54.00	170.00	670.72	692.71	700.01	709.48
53.00	126.00	661.36	684.35	690.04	700.81
52.00	182.00	652.11	676.03	682.27	692.14
51.00	214.00	642.94	667.75	674.70	683.48
50.00	216.00	633.84	659.49	667.49	674.81
49.00	252.00	624.81	651.26	659.97	666.14
48.00	217.00	615.84	643.05	651.26	657.48
47.00	241.00	606.92	634.86	643.18	649.81
46.00	255.00	598.04	626.67	636.20	640.14
45.00	288.00	589.22	618.50	629.19	631.48
44.00	280.00	580.43	610.34	622.31	622.81
43.00	281.00	571.69	602.18	615.15	614.14
42.00	290.00	563.00	594.03	606.98	605.48
41.00	306.00	554.34	585.89	598.75	596.81
40.00	344.00	545.74	577.74	590.85	588.14
39.00	332.00	537.18	569.60	582.88	579.48
38.00	305.00	528.66	561.46	575.52	570.81
37.00	343.00	520.19	553.32	566.91	562.14
36.00	371.00	511.77	545.17	556.98	553.48
35.00	352.00	503.40	537.02	548.13	544.81
34.00	380.00	495.07	528.86	539.40	536.14
33.00	286.00	486.80	520.69	531.29	527.48
32.00	341.00	478.57	512.51	522.98	518.81
31.00	351.00	470.39	504.32	513.70	510.14
30.00	303.00	462.26	496.12	505.50	501.48
29.00	325.00	454.17	487.89	497.58	492.81
28.00	315.00	446.13	479.65	489.31	484.14
27.00	318.00	438.14	471.39	480.38	475.48
26.00	305.00	430.18	463.10	471.32	466.81
25.00	295.00	422.27	454.79	462.79	458.14
24.00	278.00	414.39	446.46	454.74	449.48
23.00	313.00	406.54	438.09	446.94	440.81
22.00	292.00	398.72	429.69	438.11	432.14
21.00	276.00	390.92	421.26	428.84	423.48
20.00	267.00	383.15	412.78	420.41	414.81
19.00	259.00	375.38	404.26	412.00	406.14
18.00	242.00	367.63	395.70	403.61	397.48
17.00	255.00	359.87	387.09	394.87	388.81
16.00	262.00	352.11	378.43	385.29	380.14
15.00	231.00	344.34	369.70	376.01	371.48

Table A.9 continued

Score Conversion Table for Analytical Scale of
Form 3CGR1

14.00	237.00	336.55	360.91	367.37	362.81
13.00	214.00	328.73	357.06	358.96	354.14
12.00	226.00	320.87	343.12	349.49	345.48
11.00	203.00	312.96	334.11	338.04	336.81
10.00	200.00	304.99	325.01	327.79	328.14
9.00	195.00	296.95	315.80	316.88	319.48
8.00	196.00	288.81	306.49	305.37	310.81
7.00	165.00	280.58	297.05	294.69	302.14
6.00	187.00	272.22	287.47	283.85	293.48
5.00	154.00	263.71	277.72	273.36	284.81
4.00	176.00	255.02	267.77	261.71	276.14
3.00	111.00	246.11	257.57	250.33	267.48
2.00	117.00	236.95	247.06	240.23	258.81
1.00	95.00	227.46	236.17	228.91	250.14
0.0	76.00	217.58	224.80	217.85	241.48
-1.00	58.00	207.17	212.79	207.57	232.81
-2.00	28.00	196.03	199.91	199.04	224.14
-3.00	25.00	183.52	185.77	193.02	215.48
-4.00	25.00	173.47	174.79	184.30	206.81
-5.00	10.00	164.71	165.96	171.12	198.14
-6.00	5.00	155.95	157.14	163.76	189.48
-7.00	0.0	147.19	148.32	160.84	180.81
-8.00	3.00	138.43	139.49	159.09	172.14
-9.00	3.00	129.66	130.67	151.44	163.48

Appendix B

Relative Efficiency Curves for Various Score Scales Produced by Different IRT Equating Methods on Forms 3CGR1, ZGR1, K-ZGR2, and K-ZGR3

Figure 8.1.a

Efficiency of the Verbal Section of Form [illegible] (Calibrated on February Data) Relative to the Verbal Section of Form [illegible] (Calibrated on June Data)

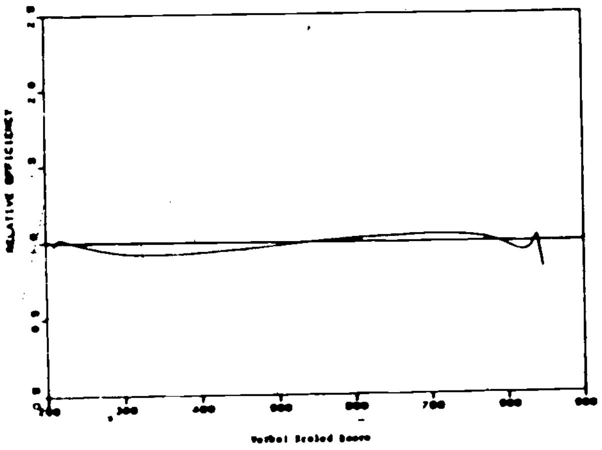


Figure 8.1.b

Efficiency of the Verbal Section of [illegible] (Based on Separate Reading Comprehension and Discrete Verbal Calibrations on February Data) Relative to the Verbal Section of Form [illegible] (Based on Separate Reading Comprehension and Discrete Verbal Calibrations on June Data)

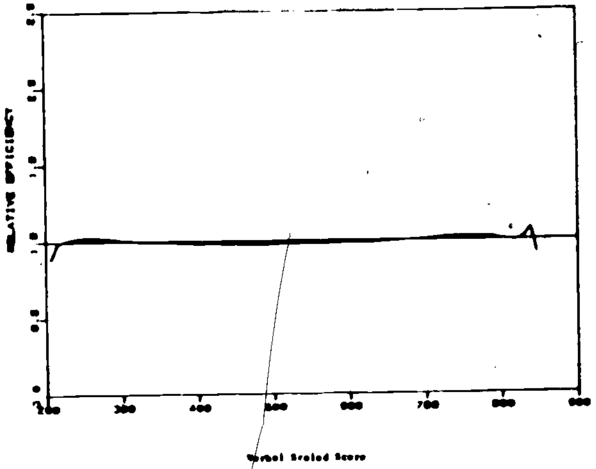


Figure 8.1.c

Efficiency of Verbal Section of [illegible] (Based on Separate Reading Comprehension and Discrete Verbal Calibrations on February Data) Relative to the Verbal Section of Form [illegible] (Calibrated on June Data)

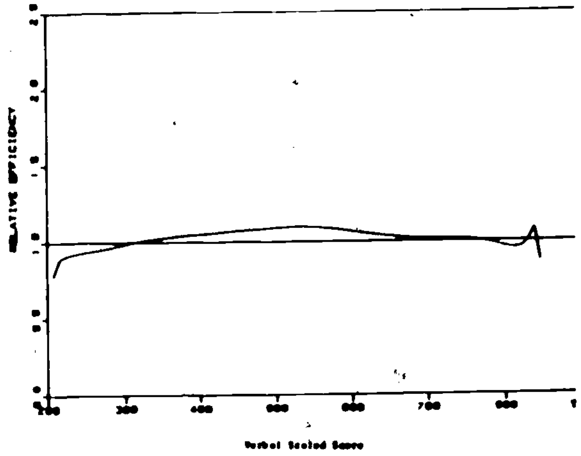


Figure 8.2.a

Efficiency of the Verbal Section of Form [illegible] Relative to the Verbal Section of Form [illegible] (Calibrated on June Data)

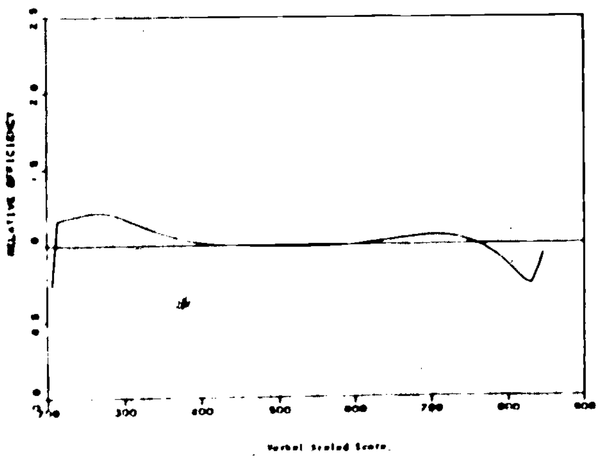


Figure 8.2.b

Efficiency of the Verbal Section of Form [illegible] (Based on Separate Reading Comprehension and Discrete Verbal Calibrations) Relative to the Verbal Section of Form [illegible] (Based on Separate Reading Comprehension and Discrete Verbal Calibrations on June Data)

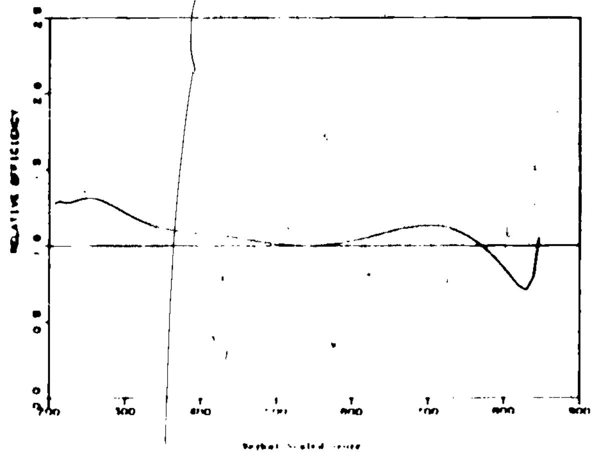


Figure 8.2.c

Efficiency of the Verbal Section of Form [illegible] (Based on Separate Reading Comprehension and Discrete Verbal Calibrations) Relative to the Verbal Section of Form [illegible] (Calibrated on June Data)

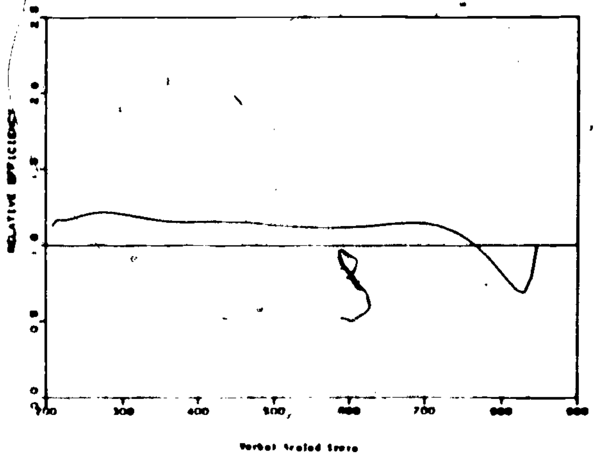


Figure B.1.a

Efficiency of the Verbal Section of Form [illegible] Relative to the Verbal Section of Form [illegible] (Calibrated on June Data)

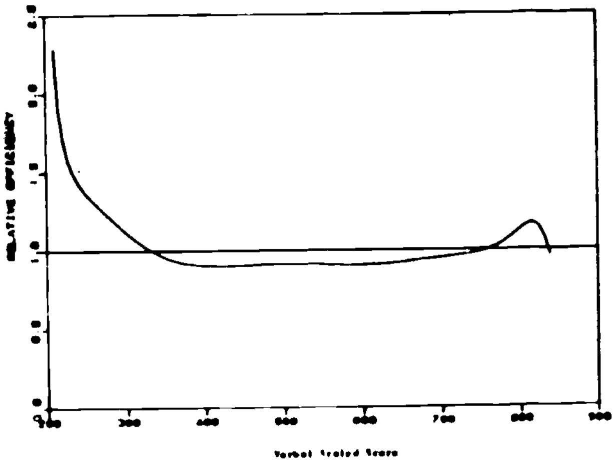


Figure B.1.b

The Efficiency of the Verbal Section of Form [illegible] (Estimated Parameters Linked to Scale by Indirect Method) Relative to the Verbal Section of Form [illegible] (Calibrated on June Data)

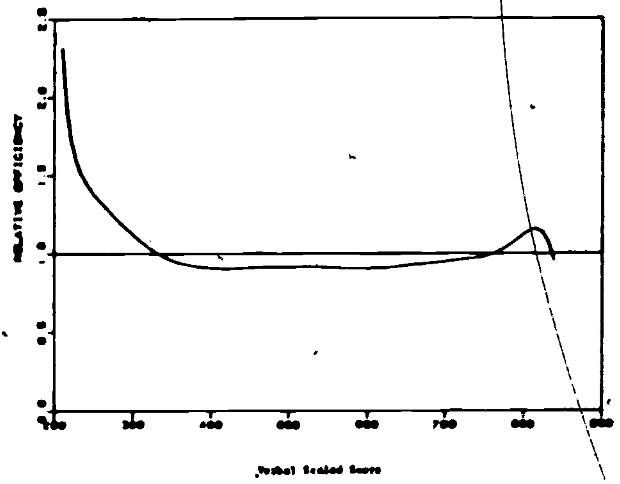


Figure B.1.c

Efficiency of the Verbal Section of Form [illegible] (Based on Separate Reading Comprehension and Discrete Verbal Calibrations) Relative to the Verbal Section of Form [illegible] (Based on Separate Reading Comprehension and Discrete Verbal Calibrations on June Data)

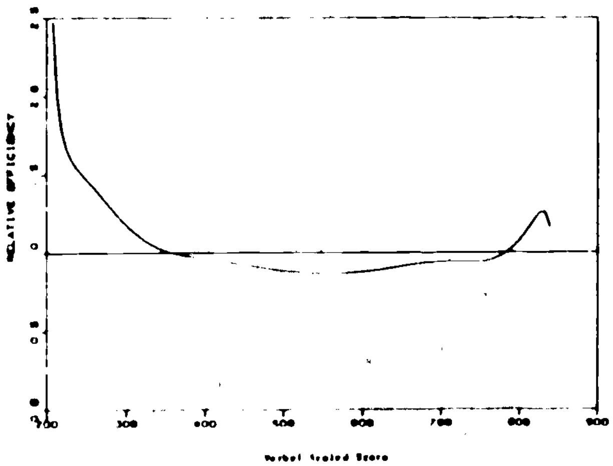


Figure B.1.d

Efficiency of the Verbal Section of Form [illegible] (Based on Separate Reading Comprehension and Discrete Verbal Calibrations) Relative to the Verbal Section of Form [illegible] (Calibrated on June Data)

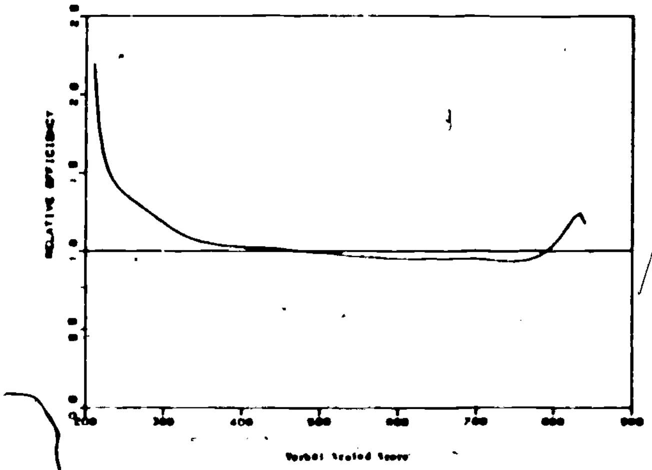


Figure B.4.a

Efficiency of the Verbal Section of Form [illegible] (Estimated Parameters Linked to Scale by Spiralling) Relative to the Verbal Section of Form [illegible] (Calibrated on June Data)

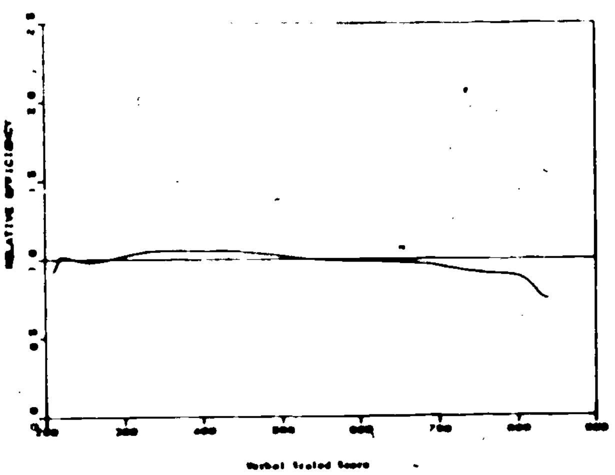


Figure B.4.b

Efficiency of the Verbal Section of Form [illegible] (Estimated Parameters Linked to Scale Using Pre-Equated Design) Relative to the Section of Form [illegible] (Calibrated on June Data)

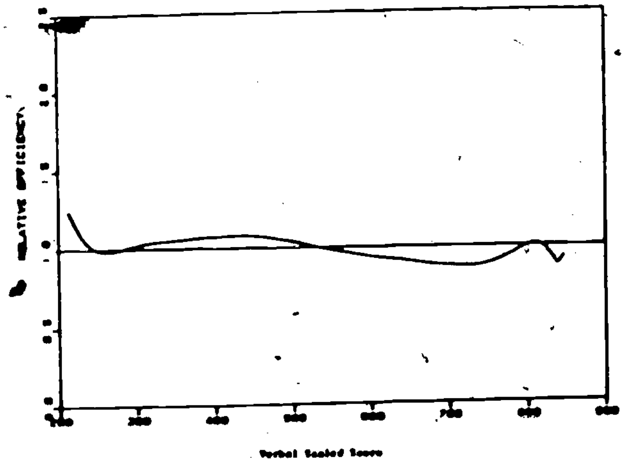


Figure B.4.c

Efficiency of the Verbal Section of Form [illegible] (Estimated Parameters Linked to Scale by Spiralling Based on Separate Reading Comprehension and Discrete Verbal Calibrations) Relative to the Verbal Section of Form [illegible] (Based on separate Reading Comprehension and Discrete Verbal Calibrations on June Data)

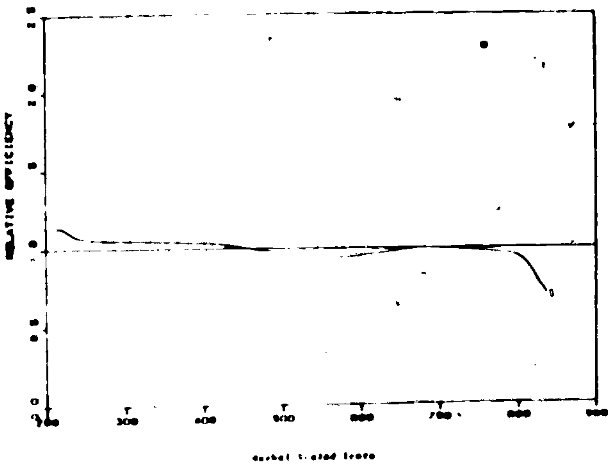


Figure B.4.d

Efficiency of the Verbal Section of Form [illegible] (Estimated Parameters Linked to Scale by Spiralling Based on Separate Reading Comprehension and Discrete Verbal Calibrations) Relative to the Verbal Section of Form [illegible] (Based on June Data)

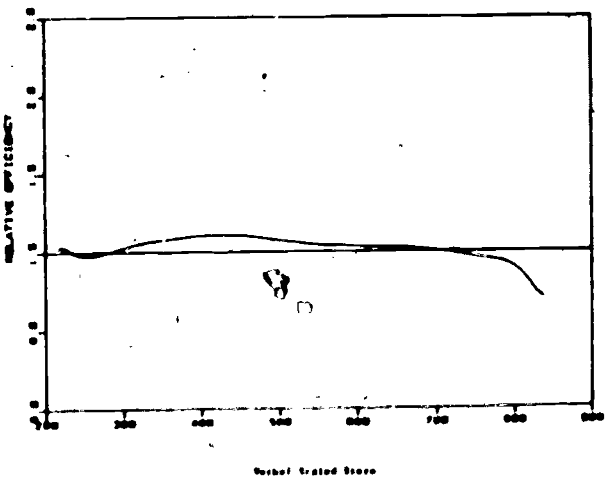


Figure B.5

Efficiency of the Quantitative Section of Form ZCR1 (Calibrated on February Data) Relative to the Quantitative Section of Form ZCR1 (Calibrated on June Data)

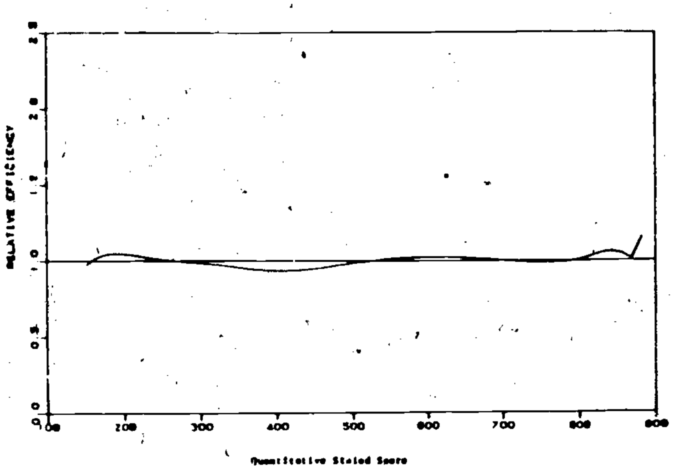


Figure B.6

Efficiency of the Quantitative Section of Form K-ZCR2 Relative to the Quantitative Section of Form ZCR1 (Calibrated on June Data)

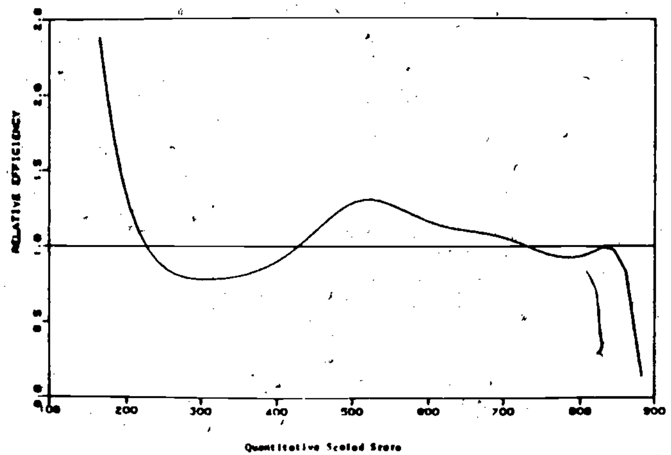


Figure B.7.a

Efficiency of the Quantitative Section of Form K-ZCR3 Relative to the Quantitative Section of Form ZCR1 (Calibrated on June Data)

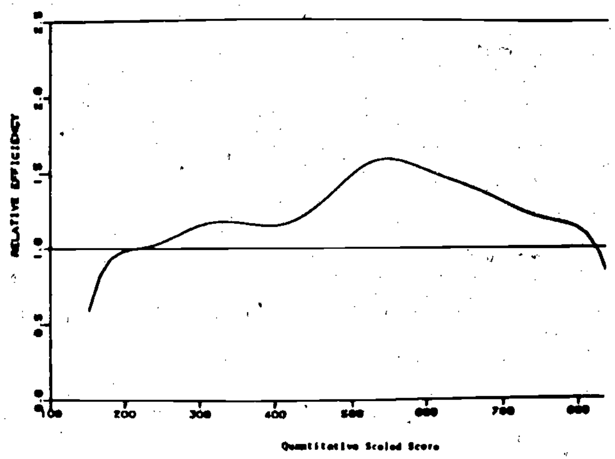


Figure B.7.b

Efficiency of the Quantitative Section of Form K-ZCR3 (Estimated Parameters Linked to Scale by Indirect Method) Relative to the Quantitative Section of Form ZCR1 (Calibrated on June Data)

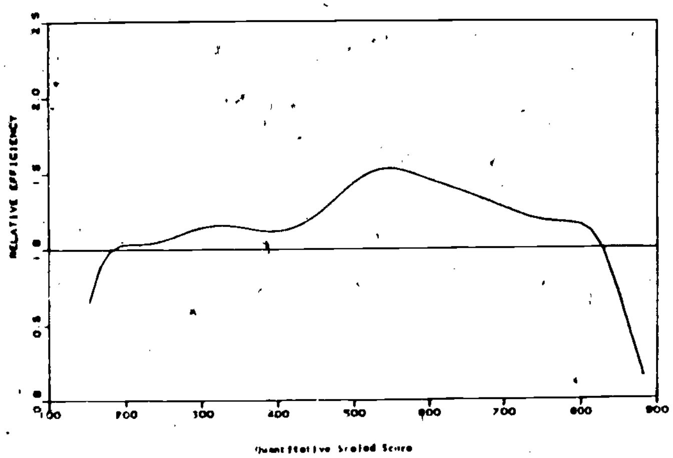


Figure B.8.a

Efficiency of the Quantitative Section of Form 3CYR1 (Estimated Parameters Linked to Scale by Spiralling) Relative to the Quantitative Section of Form ZCR1 (Calibrated on June Data)

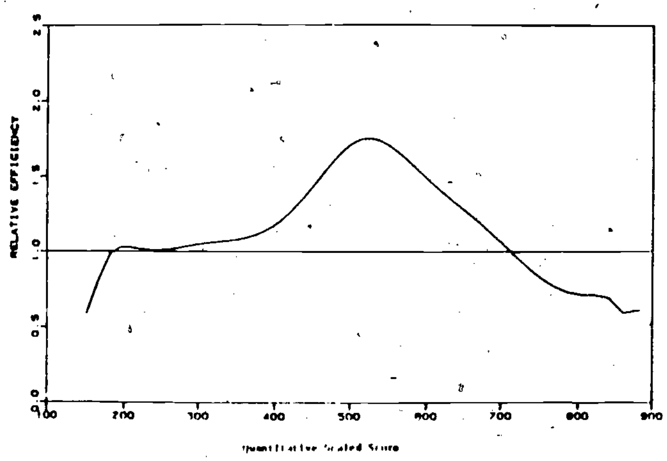


Figure B.8.b

Efficiency of the Verbal Section of Form 3CBR1 (Estimated Parameters Linked to Scale Using Pre-Equating Design) Relative to the Quantitative Section of Form ZCR1 (Calibrated on June Data)

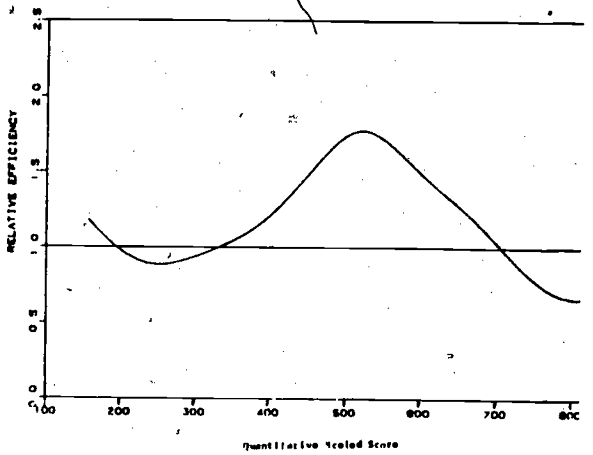


Figure B.9.a

Efficiency of the Analytical Section of Form 3CGR1 (Estimated Parameters Linked by Spiralling) Relative to the Analytical Section of Form ZGR1

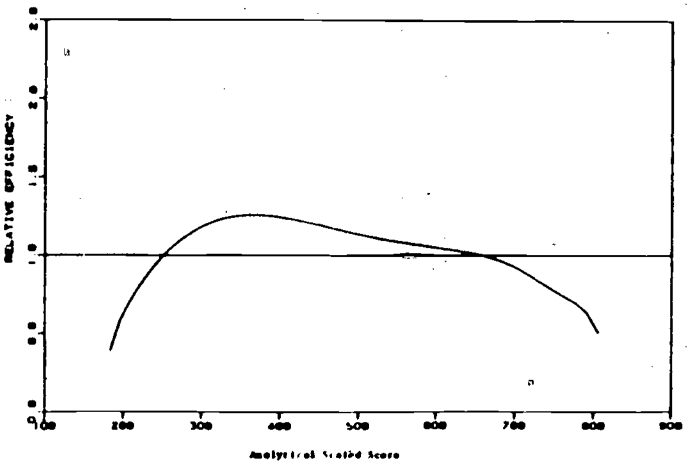
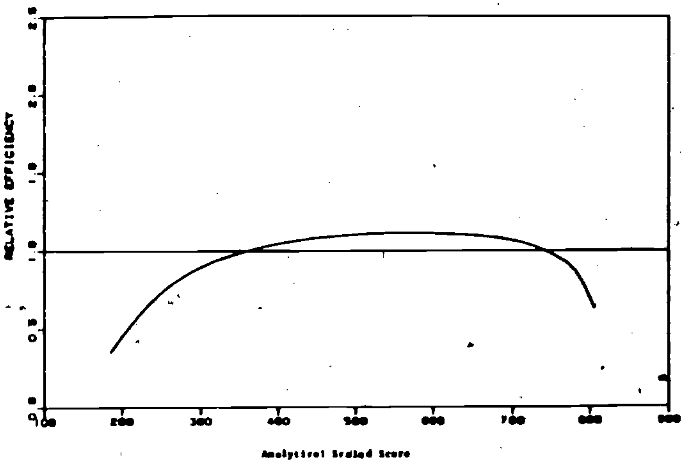


Figure B.9.b

Efficiency of the Analytical Scale of Form 3CGR1 (Estimated Parameters Linked to Scale Using Pre-Equating Design) Relative to the Analytical Section of Form ZGR1



GRE BOARD RESEARCH
REPORTS OF A TECHNICAL NATURE

Boldt, R. R. Comparison of a [redacted]an and a Least Squares Method of Educational Prediction. GREB No. 70-3P, June 1975.

Campbell, J. T. and Belcher, L. H. Word Associations of Students at Predominantly White and Predominantly Black Colleges. GREB No. 71-6P, December 1975.

Campbell, J. T. and Donlon, T. F. Relationship of the Figure Location Test to Choice of Graduate Major. GREB No. 75-7P, November 1980.

Carlson, A. B.; Reilly, R. R.; Mahoney, M. H.; and Casserly, P. L. The Development and Pilot Testing of Criterion Rating Scales. GREB No. 73-1P, October 1976.

Carlson, A. B.; Evans, F.R.; and Kuykendall, N. M. The Feasibility of Common Criterion Validity Studies of the GRE. GREB No. 71-1P, July 1974.

Donlon, T. F. An Exploratory Study of the Implications of Test Speededness. GREB No. 76-9P, March 1980.

Donlon, T. F.; Reilly, R. R.; and McKee, J. D. Development of a Test of Global vs. Articulated Thinking: The Figure Location Test. GREB No. 74-9P, June 1978.

Echternacht, G. Alternate Methods of Equating GRE Advanced Tests. GREB No. 69-2P, June 1974.

Echternacht, G. A Comparison of Various Item Option Weighting Schemes/A Note on the Variances of Empirically Derived Option Scoring Weights. GREB No. 71-17P, February 1975.

Echternacht, G. A Quick Method for Determining Test Bias. GREB No. 70-8P, July 1974.

Evans, F. R. The GRE-Q Coaching/Instruction Study. GREB No. 71-5aP, September 1977.

Fredericksen, N. and Ward, W. C. Development of Measures for the Study of Creativity. GREB No. 72-2P, June 1975.

Levine, M. V. and Drasgow, F. Appropriateness Measurement with Aptitude Test Data and Esimated Parameters. GREB No. 75-3P, March 1980.

McPeek, M.; Altman, R. A.; Wallmark, M.; and Wingersky, B. C. An Investigation of the Feasibility of Obtaining Additional Subscores on the GRE Advanced Psychology Test. GREB No. 74-4P, April 1976.

Pike, L. Implicit Guessing Strategies of GRE Aptitude Examinees Classified by Ethnic Group and Sex. GREB No. 75-10P, June 1980.

Powers, D. E.; Swinton, S.; Thayer, D.; and Yates, A. A Factor Analytic Investigation of Seven Experimental Analytical Item Types. GREB No. 77-1P, June 1978.

Powers, D. E.; Swinton, S. S.; and Carlson, A. B. A Factor Analytic Study of the GRE Aptitude Test. GREB No. 75-11P, September 1977.

Reilly, R. R. and Jackson, R. Effects of Empirical Option Weighting on Reliability and Validity of the GRE. GREB No. 71-9P, July 1974.

Reilly, R. R. Factors in Graduate Student Performance. GREB No. 71-2P, July 1974.

Rock, D. A. The Identification of Population Moderators and Their Effect on the Prediction of Doctorate Attainment. GREB No. 69-6bP, February 1975.

Rock, D. A. The "Test Chooser": A Different Approach to a Prediction Weighting Scheme. GREB No. 70-2P, November 1974.

Sharon, A. T. Test of English as a Foreign Language as a Moderator of Graduate Record Examinations Scores in the Prediction of Foreign Students' Grades in Graduate School. GREB No. 70-1P, June 1974.

Stricker, L. J. A New Index of Differential Subgroup Performance: Application to the GRE Aptitude Test. GREB No. 78-7P, June 1981.

Swinton, S. S. and Powers, D. E. A Factor Analytic Study of the Restructured GRE Aptitude Test. GREB No. 77-6P, February 1980.

Ward, W. C. A Comparison of Free-Response and Multiple-Choice Forms of Verbal Aptitude Tests. GREB No. 79-8P, January 1982.

Ward, W. C.; Frederiksen, N.; and Carlson, S. B. Construct Validity of Free-Response and Machine-Scorable Versions of a Test of Scientific Thinking. GREB No. 74-8P, November 1978.

Ward, W. C. and Frederiksen, N. A Study of the Predictive Validity of the Tests of Scientific Thinking. GREB No. 74-6P, October 1977.